VISIT…

LANZAROTE Caliente.COM

高等院校英语语言文学专业研究生系列教材

总主编　戴炜栋

语料库语言学导论

An Introduction to Corpus Linguistics

杨惠中　主编

外教社

SHANGHAI FOREIGN LANGUA

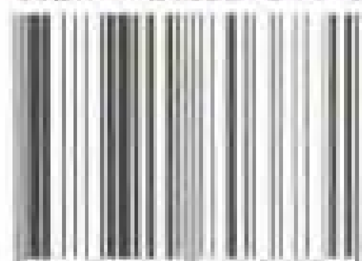
上海外语教育出版社

《高等院校英语语言文学专业研究生系列教材》是我国首套英语语言文学专业研究生教材。该系列教材汇聚了全国英语语言文学专业各领域的百余位知名专家学者，旨在集各校学术之长，优势互补，形成合力，以推动研究生教材建设，使我国英语语言文学专业研究生的培养走上一个新台阶。

本系列教材紧密结合研究生教学实际和需要，强调科学性、系统性、先进性和实用性，力求每本教材都能反映出该领域的新理论、新方法和新成果。该系列由50余种教材组成，内容涵盖语言学、语言教学、文学理论、原著选读等领域。教材内容翔实，规模宏大，体系科学，编排规范，堪称21世纪研究生教材建设的一大景观。

责任编辑 王彤福　　封面设计 龙纯立

ISBN 7-81080-373-5



定价：18.60元

高等院校英语语言文学专业研究生系列教材

总主编 戴炜栋

语料库语言学导论

An Introduction to Corpus Linguistics

主编: 杨惠中

编著: 卫乃兴 李文中

濮建忠 雷秀云



上海外语教育出版社

SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

语料库语言学导论 = An Introduction to Corpus Linguistics / 杨惠中主编. —上海:上海外语教育出版社,2002
高校英语语言文学专业研究生系列教材
ISBN 7-81080-373-5

Ⅰ.语… Ⅱ.杨… Ⅲ.词语-研究-研究生-教材 Ⅳ.H03

中国版本图书馆 CIP 数据核字(2001)第 097439 号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机），65422031（发行部）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com.cn　　http://www.sflep.com
责任编辑： 王彤福

印　　刷： 上海锦佳装璜印刷发展公司
经　　销： 新华书店上海发行所
开　　本： 890×1240　1/32　印张 13.125　字数 348 千字
版　　次： 2002 年 7 月第 1 版　2002 年 7 月第 1 次印刷
印　　数： 3 500 册

书　　号： ISBN 7-81080-373-5 / H · 150
定　　价： 18.60 元

本版图书如有印装质量问题，可向本社调换

高等院校英语语言文学专业研究生系列教材

编委会主任：戴炜栋

编　委　（以姓氏笔划为序）

王守元　山东大学
王守仁　南京大学
冯庆华　上海外国语大学
石　坚　四川大学
庄智象　上海外国语大学
朱永生　复旦大学
何兆熊　上海外国语大学
吴古华　清华大学
杜瑞清　西安外国语学院
汪榕培　大连外国语学院
陈建平　广东外语外贸大学
周敬华　厦门大学
罗选民　清华大学
姚乃强　解放军外国语学院
胡文仲　北京外国语大学
贾玉新　哈尔滨工业大学
陶　洁　北京大学
黄国文　中山大学
程爱民　南京师范大学
戴炜栋　上海外国语大学



总 序

近年来，随着我国经济的飞速发展，社会对以研究生为主体的高层次人才的需求日益增长，我国英语语言文学专业的研究生教学规模也在不断扩大。各高校在研究生培养方面，形成了各自的特色，涌现出一批学科带头人，开设出自己的强项课程。但同时我们也认识到，要使研究生教育持续健康地发展，要培养学生创新思维能力和独立研究与应用能力，必须全面系统地加强基础理论与基本方法方面的训练。而要实现这一目标，就必须有一套符合我国国情的、系统正规的英语语言文学专业研究生主干教材。

基于这一认识，我们邀请了全国英语语言文学专业各研究领域中的知名专家学者，编写了这套《英语语言文学专业研究生系列教材》，旨在集各高校之所长，优势互补，形成合力，在教材建设方面，将我国英语语言文学专业的研究生培养工作推上一个新的台阶。我们希望通过这套教材的出版，来规范我国的英语语言文学专业的研究生课程，培养出更多基础扎实、知识面广、富有开拓精神、符合社会需要的高质量研究生。

在内容上，本套系列教材覆盖了英语语言文学专业各学科的主要课程。我们总的编写指导思想是：**结合我国英语语言文学专业研究生教学的实际情况与需要，强调科学性、系统性、先进性和实用性；力求做到理论与应用相结合，介绍与研究相结合，中与外相结合，史与论相结合；广泛搜集资料，全面融会贯通，使每一本教材都能够反映出该研究领域的新理论、新方法和新成果。**本套教材的这些特点，使其有别于单纯引进的国外同类原版教材，是国外教材所不可取代的。同时，两者的作用是相辅相成的。正是由于这些特点，本套教材不仅可以作为我国英语语言文学专业研究生的主干教材，也可作为中国语言文学专业的教师与学生的参考用书。

在编写体例上，我们参照了国家标准局的有关标准以及国际上的通行做法，制定了统一的规范。每章后面，都列出了思考题和深

入阅读书目，以便启发学生思考和进一步深入研究。

教材建设是学科建设的一项重要基本建设，对学科发展有着深远的影响。我们相信，正如国外剑桥大学和牛津大学出版社出版的语言学和应用语言学教材和丛书对推动国际语言学和应用语言学的发展起了巨大作用一样，在世纪之交推出的这套系列教材，也必将大大推动我国21世纪英语语言文学专业研究生教育事业的发展，促进我国英语语言文学研究水平的提高。

戴炜栋

2000年9月



前　　言

语料库语言学的发展迄今已三十年余。从我国第一个语料库JDEST语料库在上海交通大学建成至今,也有近二十年了。语料库语言学以真实语言使用中的语言事实为基本证据,凭借现代计算机技术,采用数据驱动的实证主义研究方法,对语言、语言交际和语言学习的行为规律进行多层面和全方位的研究,从而给语言学工作者带来了一种新的理念,揭示了一种新的研究方法,开辟了一个新的研究领域。语料库语言学的数据不同于以往研究中采用的“直觉性数据”和“诱导数据”,对实际使用中的语言事实进行定性定量的描写和概括,使研究更具科学性和准确性。由数百万、数千万乃至数亿词的连续文本构成的大型语料库给语言学工作者提供了强大、丰富的数据,加上一套科学的研究手段和方法,研究人员已能够对真实语言交际的各个方面,包括词汇的、句法的、语义的、语用的、语篇的,进行深入的探讨和全面的描写;运用学习者语料库数据,研究者可对学习者的典型行为进行研究,调查不同类型和不同背景的学生的不同语言特征,发现他们语言能力发展的具体过程与阶段,探讨学生的学习策略;通过监控语料库,人们还可对语言在历史进程中的动态变化进行实时监控和观察;在语言教学领域,语料库在制定教学内容、教学目标等方面为决策者提供坚实可靠的决策依据,建立在科学依据基础上的大纲设计、教材开发和语言测试将会使教学更加有效。语料库研究还为自然语言处理、机器翻译和诸多语言工程项目提供了重要的理论启示与参照数据。近二十年来语料库语言学的发展以及所取得的丰硕成果已经引起人们愈来愈广泛的重视。语料库语言学的发展迎来了新的高潮。

通过几年的研究和探索,我的几位学生写成了这本介绍语料库语言学的小书,真诚地将其奉献给国内的语言学工作者、语言学研究生和语言教师,以期对大家了解、认识和利用语料库进行研究和教学能有所帮助。在本书的第一部分,我们对语料库语言学的一些

基本概念、语料库的种类、语料库语言学的发展概貌、语言数据的性质、一般的研究方法等问题进行了概括的描述；在第二部分里，我们讨论了语料库建设的一般原则、具体方法和程序、常用的统计手段和索引工具；第三部分是我们用语料库方法开展的部分研究。

无须讳言，本书描述和探讨的决不是语料库语言学内容的全部。由于我们的研究经验和学识水平所限，书中难免挂一漏万和不尽准确之处。因此，我们恳请广大同仁不吝指正。另一方面，语料库语言学正在蓬勃发展，尚存在许多不确定的因素，某些概念和方法会发生变化，并逐步完善。我们愿意随时就该领域里的发展与大家一道切磋。

本书各章的编著分工如下：杨惠中：第一章；卫乃兴：第三、六、八章；李文中：第二、五章；濮建忠：第四章；雷秀云：第七章；第九章由辛克莱教授撰写，由卫乃兴译成中文。全部书稿由我审阅，如有错误不妥之处，理应由我负责。

上海外语教育出版社的同志为本书的出版花费了大量精力和时间，给了我们宝贵支持。对此，我们表示诚挚的感谢。

杨惠中

2001 年 7 月



目　　录

第一部分　理论研究

第一章　语料库语言学概述 …… 3
第二章　语料库与学习者语料库 …… 33
第三章　语料库证据支持的词语搭配研究 …… 82

第二部分　分析方法与技术

第四章　语料库建设及其基本统计手段和原理 …… 131
第五章　文本索引工具及应用 …… 163

第三部分　专题研究

第六章　英语词语搭配的种类 …… 199
第七章　语料库语言学与学术英语语体研究概述 …… 237
第八章　学术英语中的语义韵研究 …… 266
第九章　千年之际展望语料库语言学 …… 296

附录

附录 1　主要术语 …… 333
附录 2　书面英语词类码 …… 343
附录 3　汉英术语对照表 …… 371
附录 4　英汉术语对照表 …… 383
附录 5　英汉人名对照表 …… 395

参考书目 …… 399

第 一 部 分

理 论 研 究

第一章　语料库语言学概述

1　一种全新的研究思路

人类科学的发展从综合向高度专门化进行分化，然后又在更高的层次上进行新的综合。早期的学术研究并没有严格的学科划分，18世纪的工业革命要求各学科有深入的发展，于是分化成数学、物理学、化学、天文学、生物学等等。每一学科都有自己的研究对象和研究方法，有自己的学科界定。随着人们对客观世界的深入理解，学科进一步细分。力学从物理学中分出来并且进一步分化为普通力学、材料力学、结构力学、量子力学等等。高度专门化使学科得到深入的发展，但缺点是各学科之间缺乏联系，不符合客观世界万事万物是相互联系的这一事实。进入20世纪中叶以后人们越来越重视跨学科的研究，实际上许多重要的进展恰恰是在学科交叉的边缘上取得的。在自然科学领域中这种例子举不胜举，如物理化学、数学物理学、生物力学等。在语言学领域，现代语言学从20世纪初诞生起一直以研究语言体系为自己的学科方向。但是语言现象涉及人类活动的一切方面，要深入理解语言的本质、理解人类的言语机制、理解人类言语活动的心理本质、理解语言与社会的关系，语言研究就必然涉及心理学、社会学、神经生理学等等，于是出现了心理语言学、社会语言学、神经生理语言学、语言哲学、语用学等众多崭新的学科；关于语言体系本身的研究也早已突破了句子的界限，在纵深方向上出现了语义学，在高于句子的层次上出现了篇章语言学。这些五彩缤纷的众多交叉学科从不同的侧面揭示语言的本质，使人们对人类的言语机制有了更多的了解，使语言学能更好地为语言实

践服务，包括语言使用、语言教学、语言工程等等。这种从综合到分化，再从分化到新的综合的过程是螺旋式进行的，新的综合是在更高层次上的综合。计算机的出现为各行各业提供了强大的研究手段。语料库语言学就是出现在语言学、计算机科学、认知语言学和应用语言学边缘上的一门新的交叉学科。

语料库语言学为语言学研究提供了一种全新的研究思路，它以真实的语言数据为研究对象，从宏观的角度对大数量的语言事实进行分析，从中寻找语言使用的规律；在语言分析方面采用概率法，以实际使用中的语言现象的出现概率为依据建立或然语法进行语法分析。语料库语言学从一个新的角度揭示自然语言的复杂性。

重视对真实语言事实的研究一直是语言学研究中的优秀传统，20 世纪 60 年代初以夸克（R. Quirk）等为主进行的“英语用法调查”（“Survey of English Usage”）就通过系统的调查建立了现代英语语料库，不过这一项艰巨的工程在当时是用手工完成的。夸克等以这一语料库为基础完成的《现代英语语法》（*A Grammar of Contemporary English*）和《英语语法大全》（*A Comprehensive Grammar of the English Language*），对现代英语进行了全面的、系统的描写，在英语语言学界产生了广泛影响。最早的计算机语料库出现在 20 世纪 60 年代初，是由纳尔逊（F. Nelson）和库切拉（H. Kuçéra）建立的 BROWN 美国英语语料库。这是一个小型语料库，总容量 100 万英语词，收集了 60 年代有代表性的美国英语语料，语料取材自各种出版物，选材时考虑到各种不同文体的平衡，选材又严格按照随机原则，因此迄今为止 BROWN 语料库仍被视为标准语料库，是现代语料库语言学的发端，对语料库语言学的发展产生了重要影响。

在 60 年代初美国语言学界占主导地位的是乔姆斯基（N. Chomsky）的转换生成语法，他认为语言学的研究对象应当是人脑的语言机制，即为什么操本族语的人有能力生成和理解无限数量的合乎语法的句子并且有能力识别不合语法的句子。语言学应当研究人脑的这种语言机制而不是语言的具体运用，他把前者叫做语言

能力（competence），把后者叫做语言运用（performance），并认为语言学的中心任务是前者而不是后者，语言学的任务就是要解释这种能力的本质。由于语言能力存在于理想的本族语者的头脑中，无法直接观察。能够直接观察的是语料，是语言运用结果，而语言运用受到各种其他因素的干扰，不能用来揭示语言的本质，因此反对研究具体语料。乔姆斯基认为，“任何自然语料都是偏颇的，对其描述不过是列举一张清单而已”（1962:59）。从 20 世纪 60 年代起转换生成语言学在语言学界占主流地位，语料库语言学因此一度进入低谷。70 年代和 80 年代，从事语料库语言学研究的学者主要集中在欧洲，包括英国、瑞典、挪威、芬兰等国家。国际语料库语言学会议也主要是人数不多的语料库语言学家的聚会。但是由于语料库语言学代表了一种崭新的研究思路，并且其研究成果在辞典编纂、语言教学、自然语言处理等方面得到了实际应用，因此很快显示出强大的生命力。语料库语言学研究于是出现了新的高潮。继 BROWN 美国英语语料库以后，由英国 Lancaster 大学、挪威 Oslo 大学和 Bergen 大学等校的学者合作建立了完全与之对应的 LOB 英国英语语料库，以后相继出现了 COBUILD 语料库、英国国家语料库等容量达到上亿英语词的大型英语语料库；其他语言的语料库以及各种专门用途语料库也相继问世，各种研究工具也不断开发。可以说 20 世纪末语料库语言学研究进入了一个新的高潮，许多国际计算语言学学术会议上，语料库语言学方面的论文要占到一半以上。这种马鞍型的发展说明了语料库语言学的强大生命力。

计算机化的语料库是语言研究的强大工具，可以服务于范围广泛的研究领域，在语言描写方面包括词汇、句法、语篇等各种层次的语言研究。语料库语言学提供了强有力的手段，用来对自然语言进行定量分析。这种从真实语料出发，对大数量的语料进行宏观研究的做法有可能产生新的研究方法、建立新的语言学模型，因此计算机语料库又是各种语言学模型的试验田，可用来检验语言学模型的正确性。

语料库语言学有着广阔的应用领域，包括词典编纂、语言教学、机器翻译、人机互动查询、自动文摘等等。20 世纪末叶以来，语料

库语言学的研究和应用正方兴未艾。

2 语言学研究的三种方法

语言学研究必然涉及语言材料,根据采集和使用语言材料的途径不同,现代语言学研究的基本方法主要有三种,即内省法(introspection)、诱导法(elicitation)和基于语料库的方法(corpus-based approach)。

2.1 内省法

内省法主要是转换生成语言学家采用的研究方法,他们以语言学家本人为资料提供人(informants),依靠自己的语感(intuitions)作为判断语言现象歧义、正误、可接受性等的依据。这是转换生成语言学家获得语料的主要来源。他们认为传统语言学、历史比较语言学、结构主义语言学只是分类之学,不能解释人类语言的本质。转换生成语言学家认为具体的语言事实是无限的、收集不完的,仅仅收集语言事实并加以分类并不能说明人类语言为什么具有创造性。既然人脑具有理解和生成无限数量的句子的能力,能够区分什么是合乎语法的,什么是不合乎语法的,语言学的任务就是要描写支配这种语言能力的规则。语言学的研究对象就不应当是具体的语料,而只能是理想的本族语者头脑中的语感。这种心灵主义的(mentalist)研究方法通常采用通过内省法自造的句子,如

Flying planes can be dangerous.
Sincerity may scare the boy.
John is eager to please.
John is easy to please.

在研究方法上，转换生成语言学家认为人的语言能力可以归结为一套公理系统，按照这套公理系统操某种语言的人可以生成和识别合乎语法的句子而不会生成和识别不合语法的句子。他们通常采用形式化的方法和专门的符号系统来揭示人的语言能力，揭示句子的句法、语义和音位结构，例如下面这一组简单的改写规则，可以生成若干合乎语法的英语句子，如 the dog chased the girl, the girl chased the dog 等：

NP → Det Noun
VP → Verb NP
VP → Verb NP PP
Det → the
Noun → girl, dog
Verb → chased

语言研究的内省法可以称作着眼于语言能力的研究方法(competence-based approach)。

2.2 诱导法

诱导法(elicitation)是一种调查方法，通过实地调查来收集人们对实际使用的语言材料的看法和人们对语言材料的心理反应，通常采用有控制的方法诱导出被试者对句子或句子中的某个成分的判断，要求被试者确定句子中有没有错误、句子的可接受程度、对句子的理解程度、以及其他类似的有关数据。采用调查方法可以使结论带有某种程度的客观性和可靠性，对某个语言事实的可接受性能得到一定的量的判断。但是语言虽然是一种社会现象，语言的使用则是一种心理现象。对某种语言形式的可接受性，人们往往没有统一的认识。个别人的语言直觉可能是不可靠的，这必然影响到调查结果的可靠性。

表 1.1 是对 76 名受过高等教育的美国人进行调查的结果，对

这些被试者来说英语是他们的母语。供调查的实验句共有五个，表中“+”号表示被试者认为实验句可以接受，“-”号表示不可接受，“?”号表示没有把握。

表 1.1　诱导法调查实例

实验句	+	?	-
1. He wants some cake.	76	0	0
2. Neither he nor they know the answer.	53	16	7
3. The old man chose his son a wife.	31	24	21
4. They aren't very loved.	4	20	52
5. A nice little car is had by me.	1	2	73

(From J. Svartvik)

由表 1.1 可以看出，对于某种语言形式的可接受性，人们往往没有统一的认识。

在研究方法上，诱导法采用问卷调查和实地调查的方法，费时费力，由于受到客观条件的限制，采用这种方法，规模不可能很大。另一方面，社会调查虽然可以提供一定的量的信息并且调查结果也有一定的客观性，但是被试者的主观判断容易受到试验者提示的干扰，例如对某个句子被试者认为不可接受，但是如果试验者多问几遍“真的不可这样说吗?”“这样说一定不行吗?”被试者就可能改变想法，或变得模棱两可，没有把握。这就会影响调查结果的有效性。

诱导法既依靠客观调查，又依靠被试者的主观判断，因此可以说是一种部分着眼于语言能力部分着眼于语言运用的研究方法(partly competence-based and partly performance-based approach)。

2.3　语料库方法

语料库语言学的研究方法以语料库为基础。所谓语料库是指

在随机采样的基础上收集的有代表性的真实语言材料的集合，是语言运用的样本。如果样本具有代表性，采样具有随机性，且样本的量又足够大，则可以认为样本就是总体的真实代表；样本具有总体的统计特征，研究语料库中的语言材料即近似于研究语言本身。由于语料库中的语言材料都是人们实际使用的语言材料，因此语料库语言学的研究结果具有可靠性和真实性。调查结果不受研究者和被试者主观判断的影响，具有客观性。此外，语料库语言学的研究还能提供有关语言使用的概率信息，这就为语言研究提供了一条全新的途径，因为语言运用也具有概率特性。

由于语料库所收集的是语言事实，也就是人们实际使用的语言，因此语料库语言学方法可以说是着眼于语言运用的研究方法（performance-based approach）。

下面列表对现代语言学研究的三种方法进行比较、分析其长处和不足。

表 1.2　现代语言学研究三种方法的比较

特　　点	语言学研究的语言数据		
	I	E	C
1. 是否简单易行、速度和费用如何	+	-	-
2. 是否受到被试态度的影响	+	+	-
3. 能否提供概率信息	-	-	+
4. 是否客观	-	+/-	+
5. 能否收集大规模数据	-	+/-	+
6. 非母语者能否使用	-	+	+
7. 能否涉及不同文体与语域	-	-	+
8. 会不会受到疲倦或犹豫不决的影响	-	-	+
9. 能否进行历时研究	-	-	+
10. 是否真实的、实际的语言运用	-	-	+

I = 内省法　E = 诱导法　C = 语料库语言学方法

（From J. Svartvik）

语言学研究中对语言事实的采集采取什么态度和方法反映了两

种对立的、截然不同的语言观。乔姆斯基认为，"今天语法理论的关键问题并不是缺乏证据，而是当今的语言理论不足以对大量的证据提供恰当的解释，这些语言理论甚至没有受到过严肃的挑战"（Chomsky 1965）。可见，在他看来语言研究的主要问题不在于如何采集语言事实。但是其他语言学家则持完全相反的观点，辛克莱（J. Sinclair）认为，"能够系统地对大数量的文本语料进行审视，使我们有可能发现一些以前从未有机会发现的语言事实"（Sinclair 1991）。

乔姆斯基在1957年的《句法结构》一书中总共分析了28个自造的例句，在1965年的《句法理论的若干问题》一书中总共分析了24个自造的例句。这样的例句是本族语者凭语感自造的，采用这类例句的优点是可以揭示本族语者头脑中语言知识的本质，揭示其抽象性和复杂性，其不足是采用这种方法只能揭示人类语言能力的某些比较狭窄的方面；另一方面，用自造例句的方法来阐述某种语言理论，这种做法本身值得怀疑。每当科学家提出一种新的理论，人们通常的期望是这种理论能够解释本来就客观存在的众多事实，而不是在提出理论的同时或事后想出一些事实，目的仅仅是为了印证自己的理论，这种做法无疑是先验论的。语料库语言学则不同，它是对已有的语言事实进行研究。这些语言事实是语言的真实运用，是客观存在，不受研究者主观判断的影响。此外，语料库语言学着重对大数量文本语料进行研究，这就为语言学研究提供了一种全新的思路和研究方法。

3 两种不同的研究风格

根据一定的语言学说对语言进行描写是语言学研究的主要任务之一。对某种语言学说进行评价通常有三条标准：解释能力、对语言材料的覆盖程度和可用性。前两者比较容易理解，而可用性指的则是这种语言学理论在语言实践中的指导作用，主要是语言教学、词典编

纂、言语运用等。随着计算机技术的发展，自然语言的计算机处理引起了人们普遍的关心和重视。这就是语言工程。人们希望计算机获得理解和生成语言的能力，这样就能够直接通过自然语言与计算机进行交流、查阅浩如烟海的书面文献等。因此，对现代语言学说的评价还可以加上第四条标准，即某种语言学理论的机器可实现性。

下面以自然语言处理为例来讨论两种不同的研究风格。任何自然语言处理系统都可以看做是一定的语言学理论模型的机器实现。现代语言学研究以主要着眼于研究语言能力还是主要着眼于研究语言运用而形成不同的学派，这种语言观的不同反映在自然语言处理研究中就形成了两种不同的方法，一种是以语言能力为基础的方法，称为人工智能方法（AI approach），一种是以语言运用为基础的方法，称为基于语料库的方法（corpus-based approach）。

3.1 人工智能方法

在自然语言处理领域里曾经占主导地位的研究方法是人工智能方法。这种方法认为自然语言处理系统就是用机器对人的智能进行模拟的系统。人的语言活动是一种由规则控制的智能活动，这种用形式化的方法表达的规则系统就是所谓的语言能力。因此，自然语言处理系统必须具有复杂的语法规则，具有逻辑推理能力，具有关于外部世界的知识。自然语言处理的基本模型是：

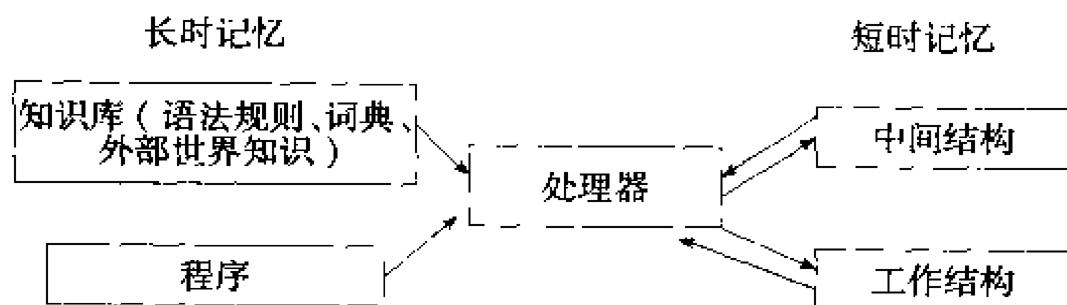


图 1.1 自然语言处理的基本模型

在实践中，人工智能方法常常从建立一个工作原型开始，这个原型只能处理一些简单的非自然的句子，例如：

1. *Whatever is linguistic is interesting.*
2. *A ticket was bought by every man.*
3. *The man with the telescope and the umbrella kicked the ball.*
4. *John and Bill went to Pisa. They delivered a paper.*
5. *Are you going to travel this summer? Yes, to Sicily.*

(R. Garside 1987)

显然，例1着眼于处理复合句，例2被动态，例3并列结构，例4代词的指代关系，例5省略形式。研究者把这些句子看做是全部可能的合格句的一个子集，环绕这一子集建立起数量有限的基本语法规则，从这些基本规则出发，通过改写规则，演绎出合乎语法的句子。采用人工智能方法的研究者希望这样的自然语言处理系统能从处理简单的句子开始，不断扩充语法规则，直至能够处理语言中的全部句子。但是他们没有意识到这种简单的句子的集合和极其纷繁复杂的自然语言真实语料之间有着难以逾越的鸿沟。

人工智能方法在自然语言处理研究领域中也取得了很大的进展，研制了不少形式语法，如ATN、WASP、LFG、GPSG等等。自然语言处理系统的句法分析器都以生成语法为基础，即首先定义各种合格的句子结构，然后用来处理自然语言中的句子。按照这些理论建立的自然语言处理系统能够在一定范围内理解自然语言的句子。例如以ATN为基础建立的LUNAR系统是一个关于月球资源的数据库，可以通过英语为用户提供查询服务，理解用户用英语提出的问题，并用英语提供查询结果。但是这种以语言能力为基础的方法有几个根本性的弱点。

1) 语言是一个开放系统，语言在本质上不是一个封闭的集合，而且语言又处在不断发展之中。把语言看做一个封闭的体系而建立起来

的自然语言处理系统,不能反映语言是不断发展的这一现实。

2) 语言是极其复杂的系统,实际使用的语言远比任何语法系统所能描写的要复杂得多,萨丕尔(E. Sapir)认为"任何语法都有漏洞"(Sapir 1921)。例如,现代英语中有一条公认的语法规则,即反身代词不可单独用作主语。但桑普森(G. Sampson)在研究 LOB 语料库时发现有这样的例句(Garside 1987):

Each side proceeds on the assumption that itself loves peace, but the other side consists of warmongers.

这里,反身代词 *itself* 单独用作了主语,而这是罗素(B. Russell)发表在《新政治家》(*New Statesman*) 的文章里的一句话。罗素是著名英国哲学家,《新政治家》又以文风严谨著称,罗素的这种用法不像是笔误,而且在同一篇文章中出现了两处,可见 *itself* 单独作主语在这里是为了达到某种修辞效果而有意使用的。

3) 设计任何自然语言处理系统的目的是处理真实的语言,而真实的语言运用不但包含规范句,而且包含大量非规范句和不规范句,即无法分析的句子,因此以语言能力学说为基础设计的自然语言处理系统,对于规范句纵然有一定的处理能力,可以按照语法规则进行推导,但是用来处理真实的语言材料时,对于其中大量的非规范句和不规范句就显得无能为力。上面提到的 LUNAR 系统,供经过培训的专家使用时成功率很高,因为查询时用的是经过选择的有限的规范句;一旦向公众开放,让没有经过训练的普通人用不受限制的英语提问,LUNAR 系统的成功率就一落千丈。

3.2 以语料库为基础的方法

以语料库为基础的方法与此不同,它的基本特征是从调查真实的语言材料出发,以处理不受限制的真实语言材料为目标。其基本方法是概率分析法,即通过对语料库中的语料进行调查,在统计分析的

基础上求出语言运用的概率信息，再反过来以概率信息为依据分析真实的语言材料。以语料库为基础的方法对语言现象进行分析时并不要求提出惟一的解，而是在若干可能的解中选择概率最高的解作为最终的解。下面以上海交通大学语言文字工程研究所黄人杰等研制的英语自动语法赋码系统 AGTS 为例来进行说明。语法赋码的任务接近于词法分析，对句子中的每个词赋予一个语法码。赋码时要提供尽可能多的词法信息、句法信息和逻辑语义信息。英语中词的用法常常是歧义的，一词多类的现象比比皆是。因此，语法赋码有一定的难度。自动语法赋码系统分三步进行赋码，第一步是给每个词赋上全部可能的候选码以及各个候选码的出现概率，如：

表 1.3　候选码实例

work	NN	VB		
		（候选码：名词、动词）		
present	JJ	VB	NN	
		（候选码：形容词、动词、名词）		
past	JJ	NN	IN	VBN
		（候选码：形容词、名词、介词、动词过去分词）		
pressure	NN	VB@		
		（候选码：名词、动词；@表示概率＜10%）		
ever	JJ	RB	VB%	
		（候选码：形容词、副词、动词；%表示概率＜1%）		

在 AGTS 系统中共采用了 131 个语法码，请参阅附录 2。

赋可能码的过程主要是查表，包括查前缀表、后缀表、例外词表和词组表。

第二步是利用上下文框架语法规则排除歧义。上下文框架语法规则的普遍形式是：

$$P_2 + P_1 + C + S_1 + S_2 \longrightarrow T \quad | \quad \overline{T}$$

其意义是:若当前词 C 出现在规则规定的特定上下文语境中(即出现在前面的词 P_2、P_1 和后随的词 S_1、S_2 之间),则必定是候选码 T 或必定不是候选码 T。请看以下规则:

$$AT + C \rightarrow \overline{VB}$$

例如在 *the work* 这一短语中,第一步查表得 *work* 的候选码,即 *work* 可能是动词(VB)也可能是名词(NN),但 *work* 作为当前词出现在冠词 *the*(AT)后面时,必定不是动词(VB),因此必定是名词(NN)。

第三步利用概率信息排除歧义。前面已经提到语言运用也是一种概率现象,设语流中有两个相邻的词,若已知前一词为冠词(AT),则后一词为名词(NN)的概率大于形容词(JJ)、远大于副词(RB)。这种信息称为相邻码渡越概率(transition probability)。通过调查现有语料库中各种语法码的分布情况,可以求出各语法码的相邻码渡越概率,其公式是:

$$\frac{F_{(T_1+T_2)}}{F_{(T_1)}}$$

即语法码 T_1 和 T_2 相邻并同时出现的频率在语法码 T_1 的出现频率中所占的比率,如:

表 1.4　相邻码渡越概率实例

AT+NN	52.64%
AT+JJ	36.99%
AT+RB	1.99%
AT+QL	1.56%
AT+JJJ	0.97%

其意义是,在冠词(AT)后紧接着出现名词(NN)的概率占冠词出现频率的 52.64%,而冠词后紧接着出现副词(RB)的概率只占

1.99%等等。

各种相邻码的相对渡越概率是通过对语料库的实际调查中得到的,因此是以经验观察为基础的。若语法码集包含131个码,则需要建立一个131×131的矩阵来记录相邻码渡越概率。在自动赋码系统中,这一概率矩阵中的概率信息可以用来排除语法码歧义,从可能的候选码中选定一个语法码,如:

表1.5 排除语法码歧义实例

there	EX
are	BER
four	CD
cutting	VBG JJ NN@
tools	NNS VBZ
.	.

其中 *there*, *are*, *four* 三个词都只有一个候选码,不存在歧义问题。但是 *cutting* 有三个候选码,*tools* 有两个候选码,因此在基数词 *four*(CD)和句号之间有六种可能的相邻码搭配,需要通过排除歧义为每个词选定一个语法码。具体做法是从概率矩阵中找出每一对相邻码的渡越概率,乘以1000,以消除小数点,其中概率最大的就是入选的语法码。

表1.6 概率计算实例

VAL(CD+VBG+NNS+.)=	1.1×32.3	×133.7=	4 745
VAL(CD+VBG+VBZ+.)=	1.1×32.3	× 30.2=	1 069
VAL(CD+JJ+NNS+.)=	41.8×34.4	×133.7=	192 249
VAL(CD+JJ+VBZ+.)=	41.8×34.4	× 30.2=	43 425
VAL(CC+NN@+NNS+.)=	26.3×(34.1×0.5)	×133.7=	59 953
VAL(CC+NN@+VBZ+.)=	26.3×(34.1×0.5)	× 30.2=	13 542

为了记录上述概率信息,需要建立一个语言数据库,这个语言

数据库的可靠性取决于语料库的大小,随着语料库的更新而更新。这一语言数据库的概率信息越可靠,自然语言处理系统的处理能力就越强,处理结果的正确程度就越高。

概率法在自动赋码方面已取得显著的成果,赋码正确率已达到98%以上。目前正在进行概率句法分析的研究,提出了或然语法(likelihood grammar)的设想。关于赋码过程的讨论,读者可参阅本书第四章。

由于语料库方法是以真实的语言材料为基础的,在这一基础上设计的自然语言处理系统的目标是处理不受限制的语料,而不是少量假想的句子。这是符合实际需要的。正因为如此,以语料库为基础的方法正在得到愈来愈多的自然语言处理研究人员的重视。

3.3　两种方法的互补

由于语言的使用深入到社会生活的各个方面,语言本身又可以从许多不同的方面加以研究,例如社会的、心理的、生理的、物理的、符号的、信息的、行为的等等。语言学家从不同的角度、不同的侧面对语言进行研究,受不同的科学思潮的影响,形成了许多不同的语言学派。这些不同的语言学派提出了不同的语言模型,他们的语言观不同,对语言本质的看法不同,解释语言事实的方法也不同。不同语言学派的研究成果有利于最终加深人们对语言的理解。以语言能力为基础的方法和以语言运用为基础的方法应当是相互补充的,而不应当是相互排斥的。

在机器翻译研究中,以词法分析为例,就有可能把两种方法结合起来,以提高系统的效率和分析的准确性。

由上海交通大学与法国 Grenoble 大学合作开发的 ARIANE-C 系统是用人工智能方法建立的机器翻译系统,分为 ATEF、ROBRA、TRANSF 和 SYGMOR 四个模块,分别进行词法分析、句法分析、转移、和生成处理。

下面是用 ATEF 模块对 *We assume that it is not difficult to find an example*. 这个句子进行词法分析的结果。

表 1.7　ATEF 模块的输出

WE:	STATE(1),CAT(R),SUBR(PERS),NUM(PLU), GENDER(MAS,FEM),SEM(ANIM).
ASSUME:	STATE(3),CAT(V),SUBR(VB),NUM(PLU),TENSE(PRES), SUBJR(S,INF),SEM(PROC),SEMV(PROC),VL1(N,SUB).
THAT:	STATE(1),CAT(D,R,S),SUBD(DEM),SUBR(DEM,REL), SUBS(CONJ),NUM (SIN),AMB(1).
IT:	STATE(1),CAT(R),SUBR(PERS),NUM(SIN).
IS:	STATE(3),CAT(V),SUBV(VB),NUM(SIN), TENSE(PRES),TYPE(COP).
NOT:	STATE(1),CAT(A),SUBR(ADV,NPMD,MDADV,CLAD).
DIFFICULT:	STATE(3),CAT(A),SUBA(ADJ),IMPERS(IMPED), SUBJR(INF).
TO:	STATE(1),CAT(S),SUBS(PREP),RS(AIM), UNSAFE(RS),VL1(TO).
FIND:	STATE(3),CAT(V),SUBV(VB),NUM(PLU), TENSE(PRES),JPCL(OUT),VEND(1), TYPE(ATROBA,ATROBN),SEM(PROC),SEMV(PROC), VL1(N,SUB),VL2(N).
AN:	STATE(1),CAT(D),SUBD(ART),NUM(SIN).
EXAMPLE:	STATE(3),CAT(N),SUBN(N),NUM(SIN), NEND(1),SEM(ABST). STATE(1),CAT(P).

从输出结果可以看到，ATEF 为每个词提供了大量词法信息、句法信息和部分语义信息。但这些信息都是可能的信息，而不是最终的信息，即相当于 AGTS 系统中的候选码，系统尚未作出最后决策。

用 AGTS 自动赋码系统对上述句子进行赋码，结果如下：

表 1.8　AGTS 自动赋码系统的输出

WE	PP1AS
ASSUME	VB
THAT	CS
IT	PP3
IS	BEZ
NOT	NT
DIFFICULT	JJ
TO	TO
FIND	VB
AN	AT
EXAMPLE	NN
.	.

把分析结果加以比较可以看出，两者的结果是互为补充的。例如代词 *we*，两个系统都提供了“人称代词”、“复数”等词法信息；此外，ATEF 系统还提供“有生命”、“性(阳性、阴性)”等信息；另一方面，AGTS 则提供了“主格”、“第一人称”等信息。可见，两个系统可以相互补充。又例如动词 *assume*，前者提供了可作及物动词和不及物动词的用法、作无人称动词的用法、后面可以跟宾语从句等，这些信息在随后的句法分析中可能会用到。而 AGTS 系统提供的不是可能的信息，而是确定的信息，因为在具体语境中每个词的用法都是特定的，而不是可能的。全面地列举各种可能的信息而不作选择会加重后续阶段的负担，例如 *that* 一词，ATEF 系统列举了可能的词类，即连词、关联词、指示代词和限定词，但未作抉择。而

AGTS 系统则对 *that* 的词类作了肯定的选择，即在本句中 *that*（*cs*）是连词。由于 ATEF 的分析是脱离上下文的，因此对词类作排除歧义的处理要等到句法分析阶段，即 ROBRA 阶段，才能进行，这就大大加重了 ROBRA 的负担。通常在自动翻译过程中句法分析阶段 ROBRA 所占的 CPU 时间要达到 80% 以上，仅对 *that* 一词作排除歧义处理，在 ROBRA 语件中就有 13 条规则之多。因此，如果将 ATEF 和 AGTS 两者结合起来，必将减轻后续阶段的工作负担，大大提高整个系统的效率。具体做法是先对输入的词串用 ATEF 和 AGTS 分别进行处理，把 AGTS 的输出转换成 ATEF 可以接受的数据形式，再对两者进行比较、排除歧义，作出选择决策，最后把合并后的输出馈入 ROBRA 系统进行处理。其框图如下：

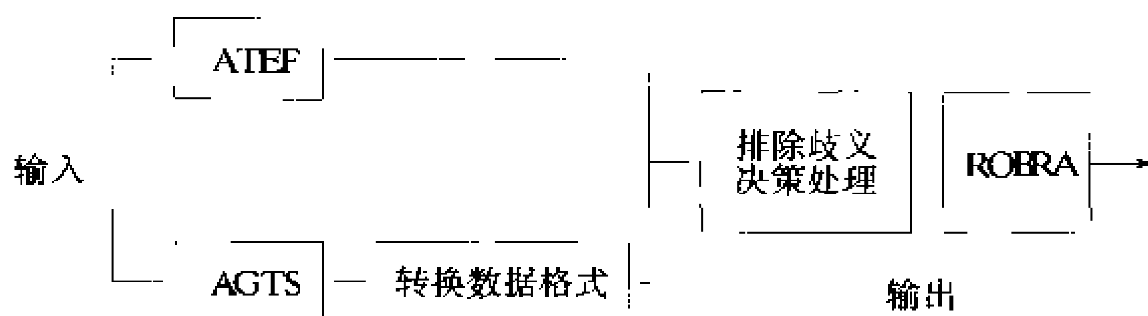


图 1.2　合并处理框图

4　语料库应用

语料库语言学的发展使语料库在语言教学、语言研究和语言工程各个领域得到了广泛的应用，下面作一简要介绍。

4.1　语言频率统计

语料库最早的应用领域之一是语言频率统计。语言频率统计

属于基础研究，如汉语中的字频统计、英语中的词汇频率统计、各种词类的出现频率统计等等。这些统计结果对语言工程、语言教学、课件开发等都有重要实践意义，这里仅以常用词教学词表的制定为例加以说明。

据统计，现代英语中的词汇总量不下数十万个，要求学生在有限的学时内掌握所有这些英语词既没有可能也没有必要。这些词出现频率不同，有的频繁出现，有的偶尔出现一次，从提高教学效率的角度来看，自然要求学生首先掌握最频繁出现的常用词。事实上，常用词表的制订一直是语言学家和心理学家关心的课题。美国语言学家、心理学家桑代克(E. Thorndike)早在20世纪40年代就着手制定常用词表；威司特(M. West)也在20世纪50年代出版了《英语通用词表》(*A General Service List of English Words*)。这些词表成了英语教学中的基础工具，是语言教学、语言测试和教材编写的基础。

教学词表的制定是一项严肃的科研工作，在计算机出现以前，常用词表的制定只能采用手工的方法，逐词抄卡片分类登记，其工作量之浩大，可想而知。计算机的出现为常用词频率统计提供了强大的科学手段。80年代初我国也完成了一项大学英语教学词表的研究工作。随着我国的改革开放，我国大学生迫切要求掌握英语，英语成了大学里的主课，但学时有限。为了在有限学时里让学生的英语达到教学大纲规定的教学目标，当时决定开展常用词教学词表的研究。制定教学词表当然必须把最需要的词收入容量有限的常用词表，但是对于哪些词应当入选，哪些词不应当入选，很容易引起争论。由于各人经历不同、阅读趣味不同、背景不同，争论可能缺乏依据，难以取得一致意见。为此在制定教学词表时确定了“定量分析为主，定性分析为辅”的原则。所谓定量分析就是通过调查研究在频率统计的基础上对常用词进行量化分析，为决策提供依据。为此专门建立了JDEST学术英语语料库，其构成如下：

表 1.9 JDEST 学术英语语料库构成

专业学科领域				
文科	~25%	教育学	语言学	历史学
		社会学	文学理论	经济学
		管理科学		
理科	~25%	物理	化学	地理学
		心理学	天文学	
工科	~30%	计算机科学与应用		机械制造
		铸造	电器工程	民用建筑
		船舶工业	航空工业	原子能
		电子工程	航天	通讯
		焊接	地质	自动化
		石油工业		
生物医学	~20%	生物学	环境科学	基础医学
		预防医学	临床医学	

为了满足学生未来的阅读需要,对于所采集的语料的文类 (genre) 也作了一定的规定,达到以下比例:

表 1.10 JDEST 学术英语语料库文类构成

JDEST 学术英语语料库文类	
期刊	25%
教科书	25%
专著	15%
论文	10%
科普读物	10%
文摘	5%
手册、书评、新闻报道等	10%

专业学科领域分布和文类分布是通过需求分析确定的，目的是使语料库具有代表性，能够反映各专业大学生未来的阅读需要。代表性是语料库建设的首要原则。在专业学科领域和文类确定以后，具体语料的采集则严格遵照随机采样的原则，以避免人工选择的主观干预。每篇语料定为不少于 2000 词的连续文本，以反映真实语言的语篇特征。学术英语语料库的容量目前达到 400 万英语词。

语料库建成后首先经过削尾处理（lemmatization），把同一个单词的不同词形的出现频率归并到根词中，如 *do*、*does*、*did*、*done*、*doing*，其出现频率都归并到根词 *do* 中，然后进行词频统计，生成频率词表。选词的主要统计特征依据的是频率、覆盖率和分布率。所谓频率是指某个词在语料中出现的频数。词的出现频数显然受所选语料的影响，有些词在某一语料中出现频数很高，换一批语料，就不一定很高。但当语料库总容量足够大时，词的出现频率就相对稳定，此时某词的频率可以看做该词在语言运用中的概率。频率词表就是将词按频率大小由高至低进行排序的词表。覆盖率指的是从频率词表上按频率次序选取的一定数量的单词，确定它们在全部语料中所占的百分率。例如频率词表上从第 1 个到第 1000 个单词，把它们的概率进行累加，就是这 1000 个词的覆盖率。派尔马（Palma）用人工计算，前 1000 个英语常用词的覆盖率为 85%，前 3000 个英语常用词的覆盖率为 92%。覆盖率是一个很重要的概念，词频统计的大量结果显示，（以 BROWN 语料库为例），前 1000 个常用英语词形的覆盖率为 68.86%，前 3000 个常用词形为 80.56%，前 5000 个常用词形为 86.20%。请注意，这里指的是词形，如果经过削尾处理，以根词为单位来计算，则前 5000 个英语常用词覆盖率可达 98%左右。十分明显，掌握这些常用词在语言教学中应给予优先地位。因为即使全部掌握了其他几十万个非常用词，对任何阅读材料来说，覆盖率也不会达到 2%。常用词不但覆盖率高，而且必定是最积极的词，无疑应成为语言教学的重点。

JDEST 学术英语语料库的频率词表实样如下：

表 1.11 JDEST 学术英语语料库频率词表实样

	频率	篇章分布率	专业分布率	大类分布率	选词指数	概率
the	274 087	6314	32	4	1000	0.076 533
of	155 584	6314	32	4	977.388	0.043 443
be	155 491	6314	32	4	977.364 1	0.043 417
and	98 596	6312	32	4	959.161 8	0.027 531
a	98 558	6315	32	4	959.162 7	0.027 52
in	86 648	6310	32	4	953.992 6	0.024 195
to	86 478	6310	32	4	953.914 2	0.024 147
that	34 946	6005	32	4	916.033 6	0.009 758
for	33 713	6125	32	4	915.277 5	0.009 414
as	27 841	5986	32	4	906.848 6	0.007 774
have	27 418	5860	32	4	905.507 9	0.007 656
with	25 865	5945	32	4	903.673 2	0.007 222
by	25 690	5934	32	4	903.338 6	0.007 173
this	22 118	5757	32	4	896.322 1	0.006 176
it	21 288	5533	32	4	893.434 2	0.005 944

频率词表上的词数和覆盖率的关系不是线性关系，其关系示意图如下：

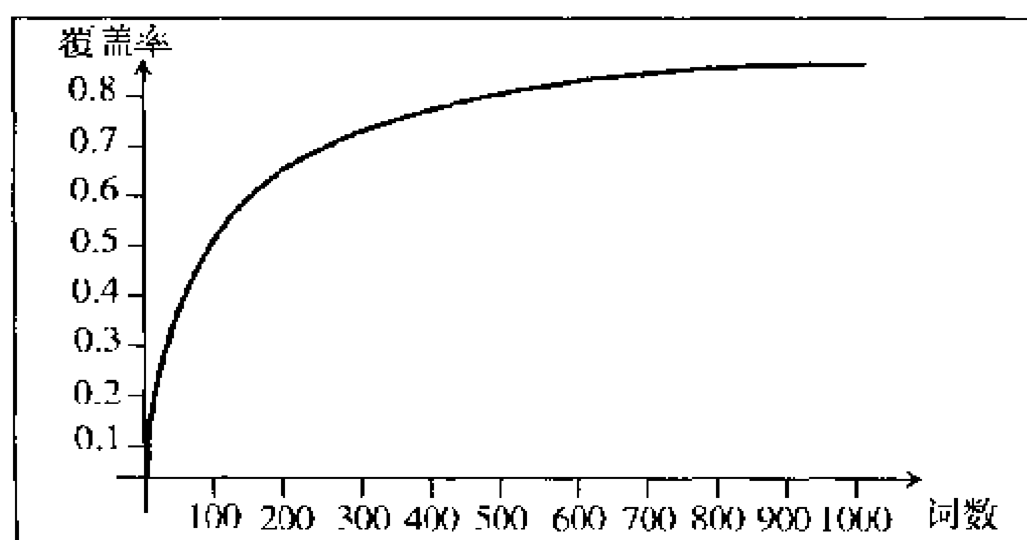


图 1.3 频率词表上词数与覆盖率关系

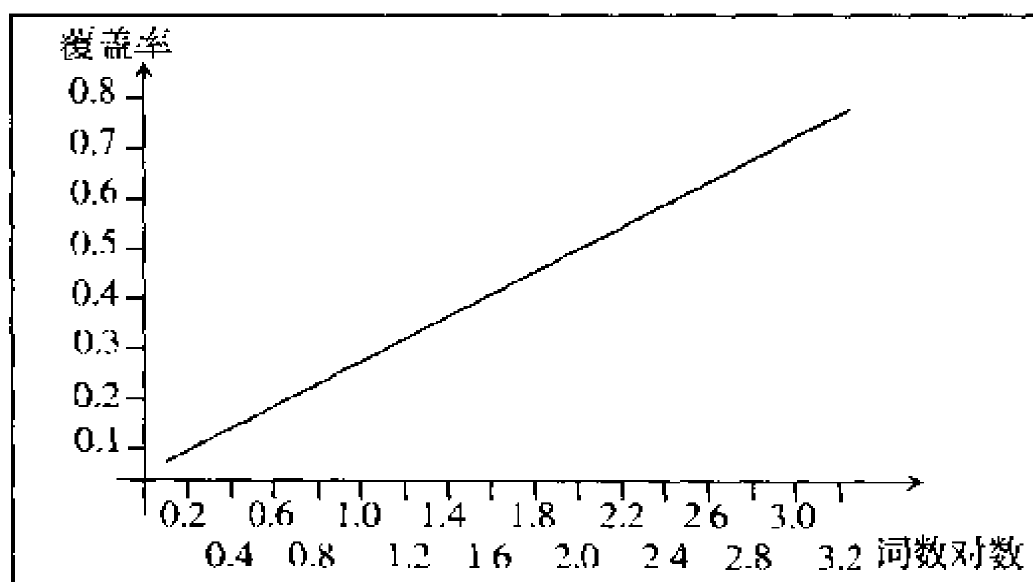


图 1.4　频率词表上词数对数与覆盖率关系

另一个重要的概念是分布率。有些词在不同专业学科领域和不同语体的文本中出现的频率都很高；另一些词则不然，只出现在某些语篇或一定的专业学科领域中。例如下面四个频率相同的词在不同语篇和专业领域中的出现频率就大不相同。

表 1.12　频数相同的 4 个词实例

单词	频率	篇章分布率	专业分布率	大类分布率	选词指数
besides	127	120	31	4	555.056 31
wait	127	99	28	4	540.713 58
carbonate	127	62	12	3	465.316 71
annulus	127	28	5	2	391.345 56

通过分析词的统计特征，可以发现三类词。一类是功能词，不但频率极高，而且分布率也极高。功能词主要是介词、代词、连词等，是一个封闭的集合。另一类词是专业术语。这类词在某个专业领域中

频率极高，但跨一个领域出现频率就可能很低，甚至根本不出现，因此分布率极低。第三类词介于两者之间，称为次高频词。这类词频率很高，但不如前两者高，而分布率却比第三类词高得多。显然功能词和次高频词是基础阶段英语教学的重点，而专业术语是随着专业课程的学习逐步自然积累的，不是基础阶段语言教学的重点。

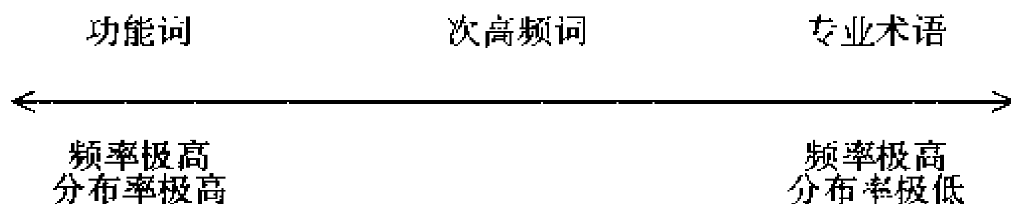


图 1.5　三类词的统计特征

为了综合考虑频率、分布率和覆盖率因素，可以采用下列选词公式：

$$I = \{\alpha \log F + \beta \log D_t + \gamma (D_s - 1)\}^{0.5} \times 1000$$

其中：

I:	选词指数
F:	频率
D_t:	篇章分布率
D_S:	专业分布率
α, β, γ:	经验数据

上述公式中 α、β、γ 都是经验数据。

通过选词指数将常用词排序，根据教学词表容量，选择位于最前面的词，例如 5000 词，必定具有最大的覆盖率，这就为科学选词提供了客观的量化依据。

根据"定量分析为主，定性分析为辅"的原则，经过上述定量分析确定的词表还需要通过定性分析进一步筛选，筛选的依据有三个，即社会学标准、语言教学标准和语言学标准。所谓社会学标准是指语言教学要适合我国国情，由于学术英语语料全部取自

国外出版的各专业领域学术原著，不可能反映我国国情，因此应当适当增加一些反映我国国情的词汇。同理还应增加一些适合课堂语言教学需要的词汇。所谓语言学标准，主要是指词的搭配能力、派生能力、联想能力、词的可用性、语义多寡等。通过定性分析对词表中的词作进一步的增删。但通过定性分析所调整的词条应是常用词教学词表中的少数，制定教学词表的主要依据是基于语料库的词频统计结果。

除了制定教学词表外，基于语料库的频率统计分析技术还可以用于术语自动识别、语体分析、自动文摘等方面，这里不进行详细讨论。关于语言频率统计的有关方法，读者可参阅本书第四章。

4.2　词典编纂

词典是使用中的语言的记录，从词条的选择、义项的确定、词义的阐释、例句的选用，无不反映编纂者的语言观，辛克莱教授在20世纪70年代带头建立了COBUILD语料库，采用词语索引技术对海量语料进行大规模调查，从此开创了现代词典编纂的先河。辛克莱教授强调的实证(evidence)是保证词典中的例句自然真实的必要条件。在COBUILD词典中，每个词条不但有频率信息，而且义项的取舍和排列次序，无不以大型语料库的实际统计结果为依据，而且每个例句必采自收集在语料库中的实际使用的语言事实。自从COBUILD词典问世以后，建立语料库已经是当代编纂原创性词典的必要条件。

4.3　词汇搭配研究

词的搭配是语言的固有特征之一，词的搭配往往是不能跨语言的，在一种语言中可接受的搭配，在另一种语言中不一定可接受。学习一门外语，在掌握了基础语法进入高级阶段以后，语言是不是地道，可以说主要看搭配。词的搭配又受到词义、用法、文化、习惯

等多种因素的影响。搭配研究本来主要靠语言学家的语感,大容量语料库的问世为搭配研究提供了客观的量化分析的依据,使词汇搭配研究更科学、更全面。这方面的研究可参看本书第三章和第六章。

4.4 语言教学

由于语料库是语言事实的采样,这就为语言教学提供了真实的语言材料。学生可以自己到语料库中查询词的用法、词的搭配、词义的细微差别等等,这就是所谓的数据驱动学习(data-driven learning)。数据驱动学习不但为学生提供真实的语境,而且为学生提供了一种探索语言的手段,学生可以像语言学家研究语言一样对语言进行主动的探索,这在写作教学中可以收到很好的效果。有关数据驱动学习的讨论,读者可参阅本书第二章。

4.5 自然语言处理

语料库语言学为自然语言处理提供了概率方法,为自然语言处理研究开辟了新的途径,由于概率是语言运用的固有特征,因此基于概率分析的自然语言处理系统对不受限制的极其复杂的真实语料的处理,成功率要高得多,而且系统结实(robust),在遇到自然语言中大量存在的不规范句或部分规范句时系统不会中断。语料库语言学方法在语音识别系统中早就得到了广泛的应用,在机器翻译和其他自然语言处理系统中也愈来愈得到研究者的重视。此外,由于语料库是真实语言材料的集合,因此也为自然语言处理系统中的语件开发提供了坚实的基础,可以大大提高整个系统的可靠性和效率。

以上关于语料库的应用只是举例性质,实际上语料库语言学的应用领域正在日益扩大。

5 语料库开发

5.1 语料库类型

根据语料库的用途和性质，迄今为止已建成使用的语料库有以下类型：

通用语料库：如 BROWN 语料库、LOB 语料库。前者是当代美国英语语料库，后者在构成上完全与之对应，不过取材自当代英国英语语料。

专用语料库：如Helsinki Corpus of Historical English，用于研究古英语；JDEST 学术英语语料库，用于研究学术英语。

监控语料库：称为 monitor corpus，用于观察现代英语的变迁，如 COBUILD 语料库。

口语语料库：如 the London-Lund Corpus、the Corpus of Spoken American English。口语语料库是研究口语特征的重要工具，如语音语调的规律，其研究成果在语声合成中有重要应用。口语语料库的建设涉及口语真实语料的采集及语音转录，工作量极大。

学生英语语料库：如 Chinese Learner English Corpus，将各种程度的学生在学习英语过程中的言语输出输入计算机，建立学生英语语料库。一般认为学习一门外语是一个渐进的过程，从接近自己的母语到逐渐接近目的语，称为中间语，学生英语语料库对于研究中间语的性质及找出学生易犯的错误，从而提高学习效率是一个重要的工具。

平行语料库：称为 parallel corpura，把两种语言中完全对应

的文本(如法律文件)输入计算机,通过分析对比找出两者对应关系,可用于机器翻译研究。

5.2 语料库规模

早期的语料库,如 BROWN 语料库和 LOB 语料库,都只有100万词的容量,称为标准语料库。这样的规模对于一般的研究来说是可以接受的,但是如果要研究搭配关系,显然容量太小,因为搭配是以两个或两个以上的词的共现为基础的,如果其中一个词出现频率很高,另一个词由于受库容量限制,出现频率不高,则两者共现缺乏显著性特征。计算机技术的发展使语料的采集和输入的困难日益减小,因此语料库的规模目前有变得愈来愈大的趋势,如 Bank of English 语料库已经达到 3.2×10^8 词次的规模。但是即使如此,百万词级的标准语料库,由于其语料的代表性、采样的随机性和各种语体比例的合理性,迄今仍然是语言研究的重要工具。

5.3 语料库加工深度

未经加工的语料库称为生语料库(raw corpus),生语料是真实使用中的语言采样的集合。生语料可以用于进行频率统计和 KWIC 词语索引查询,也可用于进行各种定性定量的研究。如果对生语料进行一定的标注,则可以提供各种信息,这一过程称为赋码,根据赋码过程所提供的信息,赋码过程所赋的可以是语法码、语义码、言语失误码等等。语法码除包括词类信息外,一般还尽量增加词法信息、句法信息、逻辑语义信息等。语法码的赋码目前已能自动进行,对不受限制的语料进行自动赋码的正确度可达到98%以上。语法码一般是在词层次上进行的,语法码赋码完成后可以加工成短语码、句法码等等。语义码和言语失误码的赋码目前基本上是人工赋码或机助人工赋码,因此如何保证赋码的正确性和一致性是赋码工作者的重要任务。

语料库建设中语料库的容量、加工深度及语料库结构三者的关

系如下图所示：

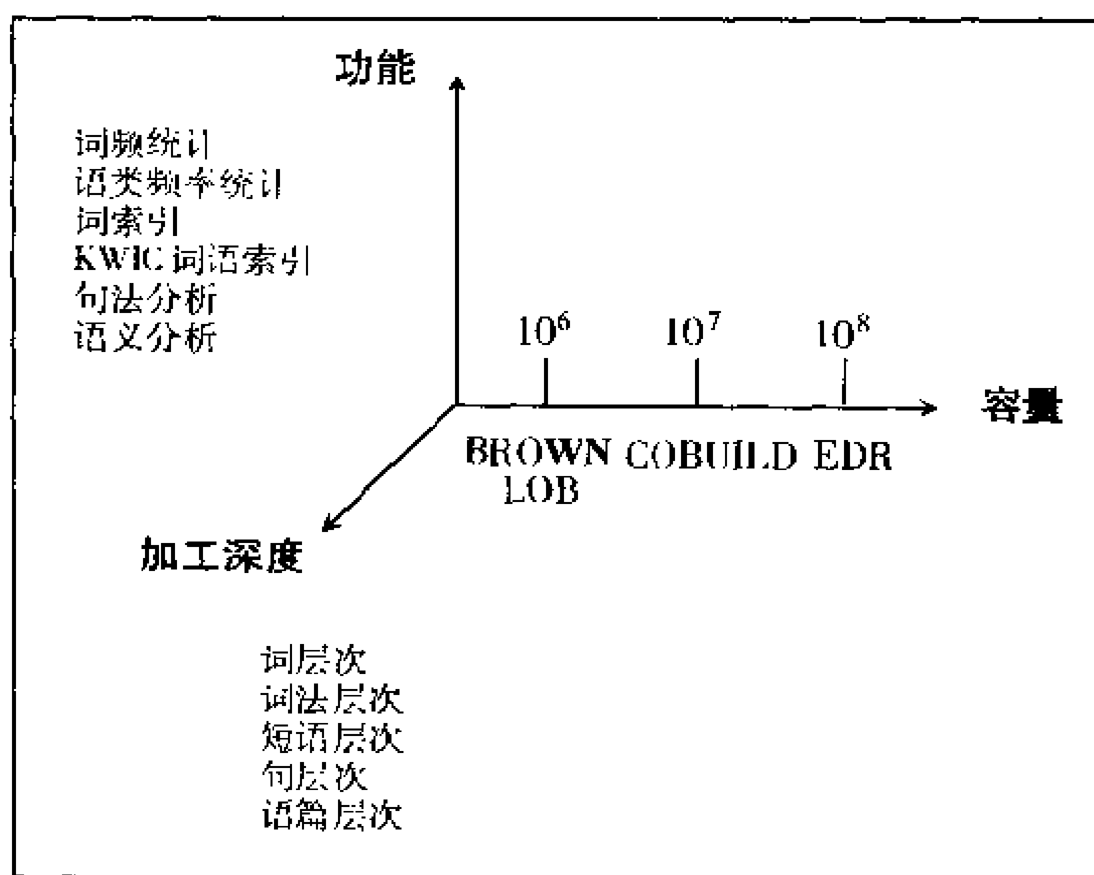


图 1.6 语料库容量、加工深度及结构三者的关系

5.4 语料库加工路径

语料库中的语言数据有些可以完全由机器自动处理，如各种频率统计、词语索引、各种查询等等；有些则要采用人机互助的加工方法，如各种赋码过程，即使是语法码的自动赋码系统，在开发的早期也必然涉及大量的人工分析、人工校对、检查等等，在这些基础工作完成以后才可能逐步过渡到自动处理。

5.5 语言数据库

语言使用也是一种概率现象，因此语言的概率信息是极为重要

的语言信息，包括词的出现概率、词的搭配概率、各种码的出现概率、相邻码的渡越概率、共现概率等等，这些概率信息不但可以提高自然语言处理系统的效率，也是各种自然语言处理系统成败的关键。语言的概率信息是对语料库进行统计分析的结果，要保证统计结果的精度，语料库容量必须足够大。但是语言是一个开放系统，而且语言又处在不断发展的过程中，因此所谓语料库容量足够大只是一个相对的概念。求得精确的语言概率信息是一个渐进的过程，首先对现有语料进行统计分析，得到各种初步的概率信息，然后用来处理新的语料(A)，这时可能会发现语言数据不够精确，需要扩大和更新(B)，这些扩大和更新了的语言数据可用来进一步提高自然语言处理系统的能力(C)，可以用来进一步处理新的语料，这样就形成了一个良性循环。这些不断更新的语言数据应当有专门的地方存放，这就是语言数据库。语言数据库记录的是语言的统计特征和概率信息，它是相对于语料库而独立存在的。

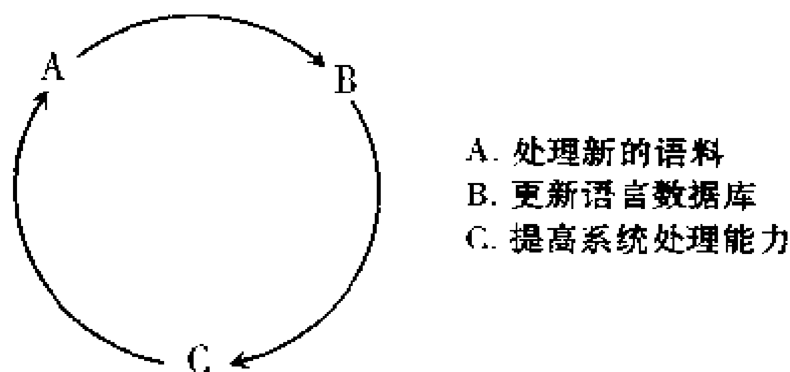


图 1.7　语言数据库信息更新

5.6　语料库语言学工具

语料库语言学发展到今天已经开发了不少通用的工具，用于词语索引、搭配研究、频率统计、削尾处理等等，详细情况读者可参阅本书第五章。

第二章　语料库与学习者语料库

1　绪论

语料库(corpus 或 corpora,corpuses[复])是指按照一定的语言学原则,运用随机抽样方法,收集自然出现的连续的语言运用文本或话语片段而建成的具有一定容量的大型电子文库。从其本质上讲,语料库实际上是通过对自然语言运用的随机抽样,以一定大小的语言样本代表某一研究中所确定的语言运用总体。作为一种语言学研究方法,语料库及词语索引(concordancing)早在 18 世纪就在欧洲得到了应用。当时的语料库大多以手工方法收集,其索引和分析过程也都是通过手工进行的,极为耗时费力。到了 19 世纪,语料库方法在语言学研究中继续得到运用,基于语料库的研究主要集中在词典编纂和语法研究方面。20 世纪 50 年代,基于语料库的语言学研究转入低谷。80 年代中期是语料库研究的复兴时期。真正意义的现代语料库是指大型的以电子文档为主要构成的计算机语料库。

与最初及早期的语料库相比,如今的语料库在概念上已发生了极大的变化。首先,变化最为显著的是语料库的大小。60 年代初,一百万词的语料库被称作大型语料库,如今真正意义上的超大型语料库已超过 3.2 亿词。第二,语料的存储和处理手段发生了质的变化。语料库从最初的印刷品收集,发展到后来的微型胶片和磁带机存储,直到如今的光盘和硬盘存储,使得语料数据的传输和保存变得十分快捷和方便。语料库索引也从最初的手工查询和统计发展到专用索引软件。第三,语料库的种类和应用领域不断扩展。在自

然语言处理领域，语料库越来越多地应用到语音识别、语音合成以及机器翻译等应用研究中去。在语言学研究中，语料库的应用已从最初的词频统计和语法分析扩展到词语搭配研究、词典编纂、句法分析以及话语分析。经过自 60 年代至今近 40 年的发展，语料库的应用在语言学习和语言教学领域也取得了丰富的成果。第四，人们对语料库的观念也发生了极大的变化。早期的语料库研究者坚持一种机械的实证主义观点，认为语料库是语言学研究惟一可靠的依据，以此排斥直觉判断和理性分析的价值。如今，人们普遍承认语料库在检验语言学假设与数据检索方面的价值，并对实证主义和理性主义的分歧持一种折衷的观点，认为二者是可以互为补充的（参见 Leech 1987:1－15）。"语料库不再被看成是能解决所有问题的万用灵药，但作为数据的系统检索，以及检验语言学假设方面的价值被越来越多的人认可"（Leech 1991 :10）。最后，随着语料库的发展，各种与之相适应的语料库统计和分析技术越来越丰富和成熟。语料库应用工具软件从最初的简单自动赋码发展到如今的句法分析和话语分析。语料库软件的适用平台也得到扩展。早期的语料库索引软件大部分只能在 DOS 环境下运行，如今的语料库软件可以在各种平台运行，并能提供网络在线查询，如英国国家语料库 BNC、伯明翰大学的 COBUILD 以及英国翻译英语语料库 TEC 等[1]。20 世纪 60 年代初，在美国 BROWN 大学建成的 BROWN 语料库，标志着基于计算机语料库研究的开始。当时语言学界的主流思潮是乔姆斯基的理性主义，在研究方法上重推理和直觉经验。在乔姆斯基看来，语言学家的主要任务是研究人们内在的语言能力，即人们内在的语法规则系统和知识。这种系统和知识完整、无误、可靠，而人们的语言运用不过是这一系统和知识的实现。语言运用是支离残缺的和舛误多变的，不足以作为语言学研究的可靠依据。所以，BROWN 语料库的建成逆当时主流而动，以实证主义为基本理念，以人们的实际语言运用为研究对象，并认为惟一可靠的数据应该是真实自然的。任何想当然的推论、自造的例句以及未经验证的假设，或通过控制实验诱导出的数据和自我报告都

是不可靠的，不足为证的。正如辛克莱(J. Sinclair)所讥讽的那样，不可能“拿一束塑料花就去研究植物学”(Sinclair 1991:5)[2]。

纵观计算机语料库的发展，初期的语料库主要用于语法分析和语言对比(如随后建成的LOB语料库)。当语言学家试图运用语料库进行词汇研究时，语料库的容量成了一个关键因素。语言学家发现，为数很少的高频词覆盖了全部语料总量的半数以上。以杨惠中主持建成于80年代的上海交通大学科技英语语料库(JDEST)为例，其扩充后的容词量为3 686 771个词[3]，总类符量为79 979，频率最高的前50个词总频数为1 436 788，占语料库总量的39%。为提高语料库的词汇覆盖率，就必须增加语料量和语料种类。综观近30年语料库发展的历史，语料库的发展呈现两大主要特征：一是语料容量不断攀高的线性发展，另一是以语料库在类型和用途上多样化为特征的扩散性发展。语料库的用户也由原来的以语言学家、自然语言处理研究者以及词典编纂者为主流用户，发展为如今的语言教师和语言学习者逐渐成为主流用户。在语言学和应用语言学研究中，基于语料库的研究深入到语音分析(phonetic analysis)、语言识别(speech recognition)和合成(speech synthesis)、词语研究(lexical studies)、句法分析、语义分析、语用分析、话语分析、词典编纂、文化研究以及学习者语言分析等各个领域。语料库方法不仅代表着一种新的研究方法，也代表着新的思维方法、一种全新的事业(Rundell 1996:2)。可以预见，在未来的语言教学和研究中，基于语料库的思想将越来越深入人心，逐步成为语言教学诸环节如大纲设计、课程设置、教材开发、词典编纂以及课堂教学的必备手段。没有语料库数据支撑的教材开发和词典编纂将变得难以想象。

2　语料库的基本特征

语料库来自拉丁词“corpus”，原意为“汇总”、“文集”等。语料

库是“作品汇集，以及任何有关主题的文本总集”（OED），是“书面语或口语材料总集，为语言学分析提供基础”（OED）。语料库是“按照明确的语言学标准选择并排序的语言运用材料汇集，旨在用作语言的样本”（Sinclair 1986：185－203）。语料库是按照明确的设计标准，为某一具体目的而集成的大型文本库（Atkins and Clear 1992：1－16）。赫努（A. Renouf）认为，语料库是“由大量收集的书面语或口语构成，并通过计算机储存和处理，用于语言学研究的文本库”（Renouf 1987：1）。里奇指出，大量收集的机读电子文本是概率研究方法中获得“必需的频率数据”的基础，“为获得必需的频率数据，我们必须分析足量的自然英语（或其他语言）文本，以便基于观测频率（observed frequency）进行合乎实际的预测。因此，就需要依靠可机读的电子文本集，即可机读的语料库”（Leech 1987：2）。综上所述，语料库具有以下基本特征：1）语料库的设计和建设是在系统的理论语言学原则指导下进行的，语料库的开发具有明确而具体的研究目标。如20世纪60年代初的BROWN语料库的主要目的是对美国英语进行语法分析，而随后的LOB语料库基本按照BROWN语料库的设计原则收集了同年代的英国英语，目的是进行美国英语和英国英语的对比分析和语法分析。2）语料库语料的构成和取样是按照明确的语言学原则并采取随机抽样方法收集语料的，而不是简单地堆积语料。所收集的语料必须是语言运用的自然语料（naturally-occurring data）。3）语料库作为自然语言运用的样本，就必须具有代表性（representativeness）。乔姆斯基曾经批评语料库不过是试图用很小的样本代表巨量的甚至无限的实际语言材料，其结果必然存在偏差，缺乏代表性。“自然语料库存在如此严重的偏差，以至于对其所进行的描述将不过是一个词表而已”（Chomsky 1962：159）。这种批评对任何以概率统计为基础手段的研究都是有价值的（McEnery 1996：5）。但是，目前的计算机语料库可以通过控制抽样过程和语料比例关系来缩小偏差，增强语料的代表性。决定语料代表性的主要因素不外乎样本抽样的过程和语料量的大小。语料库抽样一般采取随

机抽样方法。一种做法是在抽样前首先确定抽样的范围，如BROWN语料库和LOB语料库分别以1961年全年的美国英语和英国英语出版物作为抽样范围[4]；再就是确定语料的分层结构，进行分层抽样，如把语料按文类（genre）和信道（channel，如书面语和口语等）进行分层，如图2.1所示。从各种语料的抽样比例上又可分为均衡抽样（balanced）和塔式抽样（pyramidal）。前者对各种语料按平均比例抽取，而后者对不同的语料进行不等比例抽取。

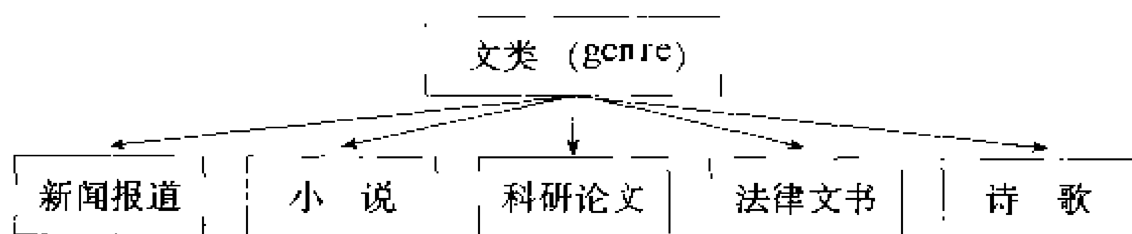


图2.1 文类分层抽样

4）语料库语料以电子文本形式储存并且是通过计算机自动处理的。巨量语料以纯文本形式存储在磁盘上，以便语料库索引软件检索和处理，也可以通过转换软件把其他格式的文件如超文本（htm或html）格式转换为纯文本。另外，语料库具有一定的容量。语料库的大小取决于语料库的设计原则和研究需求，以及建库过程中语料资源的获取难度及其他因素。计算机语料库实际上提供了一种人机交互，这种交互方式随着语料库工具的发展而逐步加强其自动化特性。里奇认为这种人机交互有以下四种渐进的模式：(1)数据检索模式：计算机以便利的形式提供数据，人进行分析；(2)共生模式：计算机提供部分经过分析的数据，人不断改善其分析系统；(3)自我组织模式：计算机分析数据并不断改善其分析系统，人提供分析系统参数及软件；(4)发现程序模式：计算机基于数据自动划分数据范畴并进行分析，人提供软件（Leech 1991:19）。

计算机自动处理包括自动词性赋码（tagging）、自动句法分析（parsing）等。其基本处理和分析过程包括以下几个步骤：

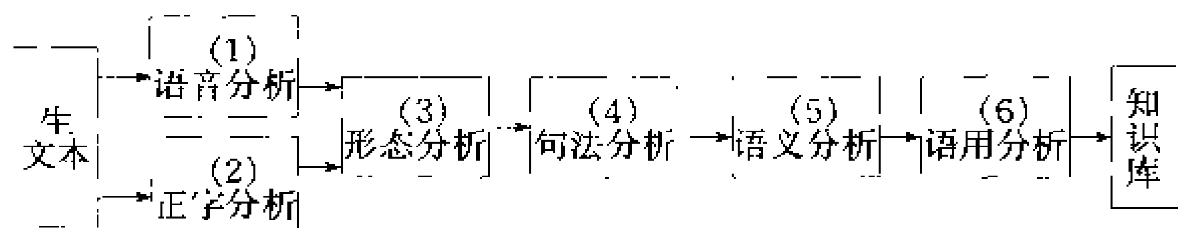


图 2.2　语料库自动分析(引自 Leech 1987:11)

语音分析(phonetic analysis)指音段分析,主要用于语音识别和语音合成。正字分析(orthographic analysis)指对文本中各种非文字符号、标点、大小写问题等进行处理和歧义消除。形态分析(morphological analysis)即词性指定和赋码。语料库自动赋码软件通过概率统计和分析,对所给句子中的每一个词指定一个或多个词性码。结果显示分为列显示和行显示两种。目前语料库自动词性赋码准确率一般在 98%以上。如句子 *I saw a saw saw a saw* 含有多个同音异义词,且各个词的词性各有不同,通过上海交大科技英语语料库赋码器(AGTS)处理后,结果相当准确(分行显示和列显示):

I	saw	a	saw	saw	a	saw.
PP1A	VB	AT	NN	VBD	AT	NN.

0001	1 —	——
0001	2 I	* PP1A
0001	2 saw	VBD
0001	2 a	AT
0001	2 saw	NN
0001	2 saw	VB
0001	2 a	AT
0001	2 saw	NN
0001	2 .	.
0001	2 —	—
0001	2 —	—

* 表中 PP1A 所代表的意义为人称代词第一人称单数;VB 代表动词原形;VBD 代表动词简单过去式;AT 代表冠词;NN 代表名词。

图 2.3　词性自动赋码

句法分析(syntactic analysis)是指句子成分切分、句法关系识别等。语义分析(semantic analysis)和语用分析对语篇进行语义指定和意义解释。5)基于语料库的研究以量化研究为基石,以概率统计为基本手段,以"数据驱动"为基本理念。其基本方法是通过对实际语言运用的抽样,确定其对语言整体的代表性,通过对样本特征的描述概括整体特征。在量化分析中,首先对特征进行分类,并统计各个特征的频率,通过建立复杂的统计模型对观测到的数据进行解释。分析结果可对研究对象总体进行概括。量化分析能够使我们发现在某一种语言或语言变体中哪些现象反映了语言的真实特征,哪些现象仅属于偶然的个例。针对某一语言变体而言,我们还可以确切地知道某一语言现象的显著性,从而确认该现象是规范的还是异常的(McEnery 1997:3)。6)语料库既是一种研究方法,又代表着一种新的研究思维,并以当代先进的计算机技术为技术手段。7)语料文本是一连续的文本或话语片断(running text or continuous stretches of discourse),而不是孤立的句子和词汇[5]。在语料库研究中,对某一搜索词的语法关系、用法以及搭配的观察是通过分析提供的语境(context)进行的。语料库索引提供的语境可分为以下几种:(1)指定跨距,即由使用者指定以搜索词为中心左右相邻的词数;(2)意元语境,即以某一意义单元结束为一微型语境,在语料库索引中意元的确定是以意义结束符号如","、";"等为标识的;(3)句子语境,即以句子终结符号如"."、"!"等为标识;(4)可扩展语境,即对搜索词所在语境可无限扩展。这对研究词汇的语法关系、词汇用法、词汇搭配、词丛(word cluster)、词汇在连续文本中呈现的范型(pattern)、以及主题词汇之间的意义关系提供了可靠而方便的途径。如 *necessarily* 一词在《新英汉词典》中作为 *necessary* 词条下该词的副词形式,定义为"必定、必然";《牛津高级学习者词典》(*Oxford Advanced Learner's Dictionary of Current English*)把它列为一个单独的词条,给出的定义为"*adv* as a necessary result; inevitably"。各种英语教科书中对该词的定义和解释也大同小

异。在上海交大科技英语语料库(以下简称 JDEST)中搜索 *necessarily* 这个词,发现该词在全库中出现 264 次,频率最大的搭配词 *not* 出现在该词左边第一个位置,观察搭配频数为 136。全库中出现 5 次以上的三词词丛有 20 组,同时含有 *not* 和 *necessarily* 的词丛有 18 组。通过索引行统计和词丛统计可以看出(见图 2.4 示例),*necessarily* 一词最典型的用法是与 *not* 搭配使用,表示含有否定意义的主观评价,意为"未必"、"不一定"[6]。如果把这个词看成是一个孤立词条并确定其定义,很难概括该词在用法中的真实行为和典型特征。

1	nd the radius of the pore is not necessarily	a function of the co
2	al to th crystal surface but not necessarily	a pure screw disloca
3	run these groups, but this is not necessarily	a good decision.
4	tish Airways did not. This is not necessarily	a criticism of BA's
5	contain quotations. This is, not necessarily	a source of ambiguity
6	in open-pool environment, is not necessarily	a prerequisite for d
7	me of calculation is not necessarily	a guide to a minimum
8	is incapacitated is not necessarily	a candidate for prot
9	whether the null findings are necessarily	accurate. Two major
10	but even that is not necessarily	achieved. The ECCE/
11	by Einstein which were not necessarily	addressed to anybody
12	problems of rural Ireland are not necessarily	all new, novel approa

图 2.4 *necessarily* 索引行示例(部分)

表 2.1 *necessarily* 词丛统计

序号	词 丛	频 数
1	is not necessarily	33
2	does not necessarily	28
3	not necessarily the	19
4	are not necessarily	17
5	but not necessarily	13
6	do not necessarily	13
7	not necessarily be	9
8	not necessarily a	8
9	though not necessarily	7
10	will not necessarily	7
11	did not necessarily	6
12	this is not necessarily	6
13	although not necessarily	5
14	necessarily the same	5
15	need not necessarily	5
16	not necessarily have	5
17	not necessarily imply	5
18	not necessarily in	5
19	not necessarily mean	5
20	not necessarily to	5

除此之外,现代计算机语料库还具有以下重要优势:1)资源优势。可获得的语料资源丰富,获得渠道方便。传统的语料库建设,语料输入工作极为浩繁,基本输入手段要靠手工键盘输入以及扫描输入。靠这种输入方式收集的语料存在大量输入错误,需进一步人工校对。如今大量的在线语料资源、光盘资料、因特网资源,包括新闻、邮件列表、电子邮件等,使得语料库的建设和扩充变得非常快捷方便。2)速度优势。早期的语料库是通过手工处理来完成分析过程的,不仅费时费力,而且误差很大,严重影响分析结果的可靠性。后来出现了在DOS环境中运行的语料库软件,提高了语料处理的自动化。但每次处理的语料量受到限制,且不易操作。另外,传统

的语料库软件大多与库本体集成开发，软件不易剥离，且适用平台少。如今，不少语料库索引软件可以在不同的操作环境中运行，且每次处理的语料量不受限制。通过专用索引软件，使得大型语料库计算机分析更加快捷。例如，只能在 DOS 环境中运行的索引分析软件 TACT2.1 每次只能处理 1 兆字节左右的语料，而如今在 WINDOWS 环境中运行的 Wordsmith Tools 可以同时处理的语料量只受计算机硬件的限制，即内存和硬盘的大小以及 CPU 速度的限制。3)精确度提高。现代语料库索引软件内嵌各种统计和检验功能，使各种统计误差更精确地体现出来(参见本书第五章)。

3 语料库的类型

从语料库这一概念的基本定义可以看出，语料库可以指任何大型的电子文本汇集。但为了进一步界定研究范围，还应把一般意义上的语料库与大型的电子文档(archives)区分开来。如上所述，语料库在语料抽样范围和文类覆盖方面都尽可能取得平衡。在收集语料时，不仅要考虑到某一文类、体裁、语域、主题类型等的抽样比例，还要考虑到建立语料库的研究目的和具体用途。与此相比，大型的电子文档更注重语料的数量和种类的丰富性。“语料库旨在代表某一预定的文类或总体。而大型文档则致力于收集任何可获得的语言材料或限定的数种文类语料，各种语言材料之间的关系要松散得多”(Edwards 1993:10)。语料库的类型可以从以下几个层面来划分：1)语言种类：可分为单语语料库、双语语料库以及多语语料库；2)语料信道：可分为书面语语料库和口语语料库；3)载体媒介：可分为印刷文本、电子文本、口语转写、影像等；4)语料状态：可分为共时语料库和历时语料库[7]；5)应用层面：可分为通用语料库和专用语料库，如学习者语料库属于专用语料库[8]；6)语料处理：可分为生语料库和标注语料库。根据编码标注的复杂程度，又可细分为

以下几种情况：(1) 不加任何处理的纯文本语料库；(2) 经过格式属性标注的语料库，如对段落、字体、字号进行标注；(3) 对识别信息进行标注，如作者、体裁、语域以及词性等标注；(4) 特殊标注，如错误赋码等。经过标注和赋码的语料库使得语料库数据分析更加系统精确，也便于对特殊数据信息的提取和处理。但是，不经任何人工介入的生语料库同样具有独特的价值。在语料库建设中，一般是保持一个干净的生语料库，而把经过赋码和句法分析的语料另存为一个子语料库或者一个独立的版本。

对那些看重语料库容量的研究者来说，语料库越大，代表性就越强。在语料库统计中，每一个在语料库中首次单独出现的词形称为类符(type)，而同一个词在语料库中出现的次数称为该词的频数，又称为该词的形符(token)。通常所说的语料库的容词量实际上是指语料库的形符总数，而每个语料库又有各自的词汇量，即类符总数。全库中所有类符及各自的频数可被统计出来，并按频数大小与字母顺序进行排序，如下表所示：

表 2.2 三个语料库中频数最高的前 10 个类符

JDEST 语料库			BROWN 语料库			LOB 语料库		
类符	频数	%	类符	频数	%	类符	频数	%
The	274 875	7.46	The	68 366	5.74	The	65 787	5.43
Of	156 177	4.24	Of	35 628	2.99	Of	34 735	2.87
And	99 029	2.69	And	28 295	2.38	And	26 872	2.22
In	87 282	2.37	To	25 593	2.15	To	26 158	2.16
To	87 157	2.36	A	23 072	1.94	A	22 225	1.83
A	79 605	2.16	In	20 931	1.76	In	20 452	1.69
Is	61 239	1.66	That	10 403	0.87	That	10 917	0.90
That	34 992	0.95	Is	9819	0.82	Is	10 430	0.86
For	33 806	0.92	Was	9745	0.82	Was	10 254	0.85
Be	32 386	0.88	He	9475	0.80	It	9705	0.80

通过语料库的类符与形符各自的比例关系，可以看出它们在语料库中的分布。语料库大小的重要性是显而易见的，这主要是由于在语料库中，占总类符量很小的常用类符的形符分布占语料库总体的比例很大，占类符量很大的不常用类符的形符分布在语料库中出现的频数却很小。就像辛克莱(1991)所指出的，一个文本中，甚至包括很长的文本，大约有一半的词汇只出现一次。为了说明这个问题，以下就 BROWN、LOB 和 JDEST 三个语料库中词汇分布进行比较：

表 2.3 三个语料库中类符/形符分布比例示例

	常用词汇		只出现 1 次词汇	
	占总类符量(%)	占总形符量(%)	占总类符量(%)	占总形符量(%)
JDEST	1.5	69.38	43	0.7
BROWN	1.9	56.32	44	2
LOB	1.9	55.55	44	1.9

从上表可以看出，在 JDEST 中，1.5%的类符的总频数占了该库总量的 69.38%，43%的类符只出现一次，其总频数占该库总量不到 1%。在 BROWN 语料库中，1.9%的类符的总频数占了该库总量的 56.32%，44%的类符只出现一次，其总频数占该库总量的 2%。在 LOB 语料库中，1.9%的类符的总频数占了该库总量的 55.55%，44%的类符只出现一次，其总频数占该库总量的 1.9%。三个语料库中，平均不到 2%的基本词汇构成了语料库总形符量的 50%以上；而 40%以上的类符只出现一次，占语料库总形符量不到 2%。这表明辛克莱的论点是非常确切的。语料库中词汇分布极不均匀。如果要研究不太常用的词汇，就需要非常大的语料库。超大型语料库以及大型电子文档在自然语言处理，包括信息检索和处理中得到广泛应用。

然而,真正意义上的现代计算机语料库并非以无限追求容量的扩张为主要目的。在语料库应用研究中,比容词量更重要的是如何根据研究者或用户的需要在语料取样中保持良好的平衡。"围绕某一可识别的文类与各种主题标准所提供的语料库材料,其构成应以用户需要为基础,即用户能够根据自己的学习和研究需要,通过汇集(语料库材料)或把语料库重新切割成各个微型语料库,获得自己的平衡和代表性"(Murison-Bowie 1993:50)。所以,并不是说把大量电子文本简单堆放在一起就建成了语料库,一个语料库的设计和建成总是为了代表某一具体领域的语言运用或满足相应的研究目的(Leech 1991:73－80)。制约语料库大小的其他因素还有口语语料库的发展与书面语语料库的不平衡问题。里奇指出,书面语语料库的语料资源由于大量增长的机读文本而易于获得,而口语体语料则由于需要语音转写,极为耗时费力,很难达到书面语语料库的同等规模(Leech 1991)。此外,里奇还指出,人类社会体系的发展往往落后于技术发展,传统的版权法使现代语料库建设难以完全自由地使用技术上能够获得的所有资源。再则,语料库建成后,由于版权限制,也不能无条件地向公众开放使用。另一与语料库发展不相适应的是语料库工具软件的开发和共享(Leech 1991:80)。目前的一些语料库工具软件如索引软件基本上只能进行一些词频统计和索引,很难满足人们对语料库进一步分析的需求。除个别软件外(如 Wordsmith Tools),大部分软件更新速度太慢,获取困难。

语料库设计标准的差异取决于以下几个因素:1)语料库是静态的还是动态的,即该语料库是有限的还是无限的。有限的语料库只收集某一固定时期的共时语言材料,语料库建成后,就不再扩充。如 BROWN 语料库只收集了 1961 年的美国英语书面语文本(共 500 篇,每篇 2000 词)。与此相比,动态的语料库保持其语料不断更新。如伯明翰大学的监控语料库就属于此类。动态语料库注重语篇的序列性和时间的连续性。监控语料库的思想首先由辛克莱提出。他认为,该语料库随着语言一同发展,"看着计算机,语言发展的全部状态就在我们眼前流过,"它不存在最后极限,从而保持对

语言发展的实时追踪(Sinclair 1991:25)。2)语料库是否经过编码标注。有些语料库已进行了语法赋码和语音标注。3)语料是完整的语篇还是片段。4)语料的选择是随意选取还是分层随机抽样。总之,语料库建设的总体趋势是语料库的容词量越来越大以及类型越来越丰富。正如赫努所指出的,"语料库的类型和大小不断扩张,各自反映语言过去和现在的不同侧面"(Renouf 1992:279)。

4 语料库的发展

对语料库本身的研究以及语料库的应用要早于 20 世纪 50 年代。基于语料库的研究方法是语言学研究中实证主义的悠久传统。西方最初的语料库索引可追溯到中世纪。以字母顺序排列伴随索引语境的词语索引的出现,最早应用在对宗教文件进行文本分析。1845 年,克拉克夫人(C. Clarke)开始对莎士比亚的 5 部戏剧作品进行了手工索引,大约花了 16 年功夫。这些可以说是现代语料库的雏形。实际上,20 世纪 50 年代以前的语料库都属于人工语料库,其建库和应用缺乏系统设计原则和有效的统计手段。20 世纪 50 年代以后,语料库语言学转入低谷,但同时第一代百万词的计算机语料库开始出现。90 年代初语料库发展急转直上,呈现出迅猛的发展势头[9]。但是,从 50 年代计算机语料库研究开始,到后来的各种语料库的兴起,其发展脉络并不连贯(Leech 1991:73 – 80)。90 年代中期,上亿词的大型语料库纷纷建成,标志着第二代语料库的开始。在不到 30 年时间里,现代计算机语料库的容量增长了一百倍。根据这一速度,里奇预言,在 21 世纪最初的 20 年内,"语料库将以千倍速度增长,达到千亿词"(1991:80)。

语料库的发展大致可分为三个阶段。第一阶段为初始阶段。这个阶段包括从 18 世纪开始到 20 世纪 50 年代计算机语料库出现前的各种手工语料库。根据肯尼迪(G. Kennedy 1998)的研究,初

期的语料库主要应用于以下五个领域:1)圣经与文学研究,如克鲁登(A.Cruden)的“钦定本圣经索引”(*Concordance of the Authorized Version of the Bible*)(1736); 2)词典编纂,如约翰生(S. Johnson)的《英语词典》(*Dictionary of the English Language*).该词典收录了15万条真实语言例证。《牛津英语词典》(*OED*)收录了414 825个词条,所引用的实例摘录达500万条,共计5000万词; 3)方言研究。4)语言教育研究。桑代克在1921年所编写的《教师手册》使用了一个450万词的语料库,制定了词汇频率表。1944年,桑代克和洛治使用了一个更大的语料库(1800万词),编写了《教师3万词手册》。5)语法研究。如叶斯珀森(O. Jesperson)(1909－1949)的《基于历史原则的现代英语语法》(*A Modern English Grammar on Historical Principles.*),克鲁辛格(E. Kruisinga)(1931－1932)的《当代英语手册》(*A Handbook of Present-Day English*),普茨默(H. Poutsma 1926－1929)的《最新现代英语语法》(*A Grammar of Late Modern English*),弗赖斯(C.C. Fries 1940)的《美国英语语法》(*American English Grammar*)等[10]。

早期的语料库还应用于儿童语言习得研究、拼写、语言教学以及比较语言学研究。自19世纪70年代至20世纪20年代,就有人利用父母对儿童语言发展所记录的日记进行收集整理,并试图依此对儿童的语言发展进行建模。这种对自然语言原始语料的应用是早期语料库的雏形。1987年,克汀(Kading)使用了一个容词量为1100万词的德语语料库,用来研究德语字母的分布频率和排列顺序。该语料库就是与现在的大型的计算机语料库相比也毫不逊色(McEnery and Wilson 1996:3)。在语言教学领域,20世纪40年代有弗赖斯和特拉弗(Traver)以及邦格斯(Bongers)等语言学家运用语料库进行教学法研究。当时的语料库主要用于外语学习者词汇表的确立。早期的语料库研究由于以下几个主要缺陷受到批评和指责:1)错误地假定自然语言的句子是有限的。2)错误地假定自然语言中的句子可以收集和列举。语料库被当作语言学研究惟

一可靠的数据源。3)否定内省研究的价值。20 世纪 50 年代以哈里斯(Harris)和希尔(Hill)为代表的美国语言学家受实证主义(positivism)和行为主义的影响,认为收集足够多的自然语言运用数据对语言学研究来说,不但是必需的,而且是自足的。直觉依据降为第二,甚至可以完全抛掉(Leech 1991:73－80)。4)受技术限制,语料量小,样本代表性差,语料处理和分析工作浩繁。按照现代语言学理论,自然语言中的句子是无限的,这不仅指词汇和句法具有的无限的可能性选择,还指自然语言自身所具有的递归性。自然语言句子的这种无限性特征,使得任何列举或穷尽收集句子的企图都是不可想象的。"解释语法的途径是描述它的规则,而不是列举其句子"(McEnery and Wilson 1997:5)。乔姆斯基认为,句法规则是有限的,而依此规则能生成的句子则是无限的。语言学家的任务是构建人们语言能力(competence)的模型,而不是收集某些语言运用(performance)。语言运用不过是人们语言能力的一些偶然的外部表征,其形式受到其他非语言能力因素的干扰。在乔姆斯基看来,语言能力集中体现了人的内在的语言知识,而语言运用不过是这种能力的"糟糕"的镜像罢了。按照这种看法,语料库从本质上讲不过是收集了一些外在的言语,属于语言运用数据,不足以为语言能力研究提供依据。关于早期语料库研究对内省研究的否定,乔姆斯基提出的关键问题是,如果我们否定人的直觉判断,不能确定人的语言能力,又如何能判定某一外部事件是否为语言事件呢?另外,完全否定内省和直觉判断的价值,人们便无法判断某一给出的语句是否合乎语法。最后,对于一个大型语料库而言,大量的数据检索和分析如果仅靠手工是极其费时费力的,而分析的过程和结果也难免存在严重误差。所以,麦克艾纳利等指出,早期的语料库语言学所要求的是当时并不具备的数据处理能力。

语料库发展的第二阶段是复兴阶段。这个阶段是以电子语料库的兴起为主要特征。从 20 世纪 50 年代到 80 年代,各种计算机语料库纷纷建成。语料库的发展以容量不断增加和种类的不断扩展为主要特征。也有人把这个时期的语料库称作第一代语料库,包

括 *BROWN Corpus*(1964)、*Lob Corpus*(1970－1978)、*London-Lund Corpus* (*LLC*, 1973)、*The Kolhapur Corpus of Indian English*(1978)、*The Wellington Corpus of Written New Zealand English*(1986)、*The Australian Corpus of English*(1986)。乔姆斯基对语料库的批评使语料库语言学研究在20世纪50年代转入低谷,但并未彻底绝迹。基于语料库的语言学研究仍时有见到。当时语言学研究的主流思潮是以乔姆斯基为代表的理性主义。但是,语料库语言学的基本方法在某些领域仍然发挥着不可替代的作用,如在语音学研究和第一语言习得研究领域,通过观察而得到的自然语言运用数据仍然是主要而可靠研究依据。这一点,连乔姆斯基本人也不得不承认。语料库发展的第三个阶段是壮大阶段。第二代超大型计算机语料库开始出现。如果说第一代电子语料库是以百万计的话,第二代语料库则以千万甚至亿计。

4.1 计算机语料库的兴起

语料库研究的复兴有以下三个方面的基础。首先是语言学基础。在英国语言学研究中,实证主义从弗思(J.R. Firth)到韩礼德(M.A.K. Halliday)至辛克莱一直传承下来。斯塔布斯(Stubbs)总结其语言学研究五大原则是:1)"语言学是一门社会科学,一门应用科学"。在这种原则下,语言本身被看成是一种社会行为,而不是独立于社会活动之外的某种客观存在。2)语言学研究所基于的数据应是"经过验证的真实的语用实例(而非直觉的或发明出的例句);语言应作为整体语篇进行研究,而不是孤立的句子或语篇片断"。实证主义的基石是对可观察的对象进行研究。作为人们外部行为的语言运用是可观察的、可靠的依据;而人们内在的语言能力是不可直接观察的,只能通过语用实例进行推断。3)"语言学的基本主题是意义研究,形式和意义不可分割的,词语与语法相互依赖"。在基于语料库的研究中,辛克莱发现,形式(form)和意义(sense)相互依存,"就像一枚硬币的两面"。抛开意义去研究形式

或抛开形式去研究意义都是不可想象的。4)“语言运用既具有常规性又具有创造性;语言在运用中传播文化”。5)“索绪尔对语言和言语的二元划分需要重新修订”(Stubbs 1996:23)。索绪尔(de Saussure)认为,人们的语言构成了一个整体(language),它包含“语言”(*langue*)和“言语”(*parole*)两个部分。换言之,在“*language*”中去掉“*parole*”便能得到“*langue*”,反之亦然(Stubbs 1996:23)。在这一理论中,作为一个抽象系统的“语言”成了凌驾于人们实际语言运用之上的某种独立的存在。在语料库语言学中,语言研究的直接对象是语言运用,只有通过对语言运用实例的大量收集和分析,才能得到语言典型特征的可靠依据。

语料库语言学复兴的另一个基础是技术发展,主要表现在以下三个方面:一,以计算机为主导的硬件技术发展。PC 机的兴起、计算机计算速度的高速增长以及存储介质(从微型胶片到磁盘、磁带机、光盘等)的开发和存储容量的急剧扩张都为语料库的发展提供了技术保障。例如上海交大科技英语语料库容词量为 360 万,占 23 兆字节,用传统的 3 英寸软盘需要 20 片。如今一张可刻写光盘容量为 650 兆,可存下 28 个相同容量的语料库。如今的大容量硬盘(如 28 千兆、50 千兆、甚至 100 千兆字节以上)以及内存(如 128 兆或 256 兆)与功能强大的索引软件结合起来,已经可以实现在单台 PC 机上同时处理多个大型语料库。二,计算机网络的发展为语料库的发展和应用提供了广阔的舞台。国际互联网的发展为语料库发展提供了有利条件。首先、大量文献和文件被转换成电子文本在网上传播,这些电子文本与网上无处不在的电子邮件和新闻讨论为语料库建设提供了取之不尽的语料源泉。过去需要花费数年才能建成的语料库,如今可能只需要几个月甚至更短时间。其次是大量语料库成为在线语料库,允许用户在网上即时使用,如伯明翰大学的 COBUILD 与朗文的 BNC 语料库等。另外就是大量网上讨论和邮件列表使得研究者和用户能够及时交流经验和观点。三,语料库索引软件的开发。如今的索引软件大多具有语料库无关性,即该软件不是为某一个语料库单独设计开发的,而是能够应用于各种类

型的甚至不同语种的语料库。如早期的 MicroConcord,TACT 以及最新的 Wordsmith Tools 等。

语料库复兴的另一个基础是需求增长。在语料库的应用领域,日益增长的用户群体和不断扩展的应用领域进一步拓展了语料库的应用价值。其应用领域可分为传统领域、扩展领域和新兴领域。传统领域包括自然语言处理、语法分析以及词典编纂等;扩展领域包括教材开发、机器翻译、语言识别和语言对比;新兴领域包括语言教学、数据驱动语言学习(Data Driven Learning,简称 DDL)、中间语对比分析(Contrastive Interlanguage Analysis,简称 CIA)、多媒体计算机辅助语言学习(Multi-media Computer Assisted Language Learning,简称 MCALL)以及在线语料库等。如今,语料库的用户群快速增长,语言教师、学生、教材开发者、语言习得研究者使用并开发自己的语料库,逐渐成为语料库的主流用户。

4.2　语料库的线性发展和增殖

兰德尔认为,现代计算机语料库发展具有两大特征,一是语料库容量的线性增长,即语料库变得越来越大,另一个就是语料库种类的不断扩展(1996:2)。表现在以下几个方面:1)多样化发展。语料库由原来的单语种发展到现代的各种双语甚至多语种语料库。另外,早期的语料库只集中在以英语为主的几种欧洲语言,而目前,世界上一些主要语言都有了自己的语料库。2)扩散性发展:多用途、多信道、多类型语料库的纷纷建成。

早在 20 世纪 60 年代初,夸克计划建立"英语用法通览"(SEU,Survey of English Usage)。同时,纳尔逊和库切拉开始建设后来闻名的 BROWN 语料库。以纳尔逊和库切拉为首建立的 BROWN 语料库为现代计算机语料库建设首开先河。BROWN 语料库收集了 1961 年全年的美国英语资料,所代表的文类十分广泛,主要构成包括信息性材料和虚构性材料。其 500 篇语言材料包含了新闻报道、社论、备忘录、宗教材料、科幻小说、侦探小说以及一

般小说故事。如今,BROWN 语料库提供全程赋码版以及部分句法赋码的版本,以供学术研究。该库所使用的赋码方案包含 82 个词性码,并使用 TAGGIT 赋码程序(Garside, Leech, & Sampson 1987)进行自动赋码。70 年代中期,斯瓦特维克和他的小组把 SEU 转换为计算机语料库,并充分结合了 BROWN 语料库和 SEU 两个语料库的特点,着手建立 Survey of Spoken English,把 SEU 口语部分转写为机读文本,最终建成了 London-Lund Corpus,即 LLC。LLC 的建成对以后的语料库研究具有重要意义。里奇认为,该语料库"至今仍是研究口语体英语无与伦比的重要资源"(1991:17)。BROWN 语料库属于静态语料库,主要用于语法分析。该语料库的建立对后来的语料库产生了深远的影响。后来的 LOB 语料库(1970)、Macquarie 语料库(1987),以及 LIMAS 现代德语语料库基本上都参照了 BROWN 语料库的建库原则。计算机技术的发展和相关软件的开发,使语料库语言学得以蓬勃发展。如今人们所谈的语料库,实际上就是指计算机语料库。为了进行美国英语和英国英语对比,在里奇(Lancaster 大学)和乔恩森(S. Johansson)(Oslo 大学)领导下,按照 BROWN 语料库的建库原则,创建了 LOB 语料库。LOB 语料库语料采集范围和所包括的文类与 BROWN 语料库保持一致,以便进行比较。

60 年代后,计算机语料库不断发展壮大,越来越深入人心,在有些领域(如自然语言处理),语料库方法实际上成了一种主流方法(Leech 1991)。随着基于语料库词典学研究的开展,人们意识到,百万词的语料库包含的大量的低频词难以满足需要。所以,80 年代语料库开发者的主要目的是建立大型语料库。在辛克莱领导下,"伯明翰英语文汇"(the Birmingham Collection of English Texts,简称 BCET)得以建立。该库的容词量达 2000 万。随后的语料库基本沿着线性的增长发展。如 Longman-Lancaster 语料库容词量达 3000 万词。20 世纪 90 年代初,容词量为 1 亿词的英国国家语料库(British National Corpus,简称为 BNC)建成;而原来的 BCET 经过扩建转换成为容量为 2 亿词的"英语文库"(Bank of

English)。该语料库是一个监控语料库,实现了语料持续动态更新。到了1996年,该语料库的容量增至3.2亿(Rundell 1996:3)。随着语料库的发展,由各种商业和学术组织负责收集的大型电子数据电子文档也纷纷建成。这些大型电子文档广泛收集了各种类型的数量可观的电子文本,以满足各种用户的不同需求。值得注意的是,发展趋势正像辛克莱所预言的那样,语料库建设的任务逐渐从语言学家手中转移到了社会科学家手中,而语言学家只专心进行语言分析。

语料库的另一个发展趋势是其类型和品种数量的高速增殖。一方面,同一语言的语料库在类型上出现分化,针对不同文类和领域建成了各种专用语料库。80年代中期,以上海交通大学杨惠中教授为首建成的JDEST就属于专用语料库。该库旨在为中国大学英语教学提供通用词汇以及技术词汇的应用信息。该库通过对个别文本词汇分布频率比较,识别出技术词汇和准技术词汇。其基本原理是,如果某些词汇在某一文本中分布频率较高而在总库中分布较低,通过比较,该部分词汇就可能是技术词汇(杨惠中1986,引自Renouf 1986:128)。这种通过语料库比较而筛选技术词汇的思想和统计方法对后来的语料库研究产生了深远的影响。JDEST语料库的建成对大学英语教学大纲词表的确定提供了可靠的量化依据,为我国大学英语教学改革发挥了重要的作用。其他专用语料库还有雷丁大学的学术英语语料库(简称RAT),以及诺丁汉大学的会话英语语料库。专用语料库发展的另一个重要趋势是学习者语料库的兴起。学习者语料库广泛收集语言学习者的论文、日记以及写作作业,成为大型的学习者语言(又称"中间语")数据库。兰德尔认为,学习者语料库在语言教学中的价值是显而易见的。它为语言教学提供了有关学习者语言运用和典型困难的可靠信息(Rundell 1996:6)。另一方面,语料库在品种和数量上呈现高速增殖的趋势。传统的仅限于少数几个欧洲语种的语料库扩展为多个语种。根据巴娄(Barlow)博士著名的语料库主页统计,目前欧洲大多数语言都有了自己的语料库。到了90年代末,世界上主要语种基本上都

建成了各自的语料库。值得一提的是，中国国家语料库工程始于90年代初，初始容量为5000万词。该语料库作为一个大型通用语料库主要用于汉语词法、句法、语义以及语用学研究（Zhou and Yu 1997:239）。如今，基于语料库的语言分析与研究已深入人心。任何对语言进行描述和分析的研究如果想得到可靠的发现，就必须以大量的客观的语言运用数据为依据。“不依赖实验数据就试图描述语言，这将被视为一种不正当的思想”（Rundell 1996:1）。

表 2.4　语料库的线性发展

序号	时间	语料库名称	创建者	容词量（百万词）
1	1967	BROWN CORPUS	Kucéra & Francis	1
2	1970's	LOB CORPUS	Geoffrey Leech & Stig Johansson	1
3	1960's ~1970's	LLC（The London-Lund Corpus）	Leech & Svartvik	0.5
4	1984~1987	SEC（The Lancaster Spoken English Corpus）	Knowles & Lawrence	0.052
5	1990	PIXI Corpora	Gavioli & Mansfield	/
6	1992	Helsinki Corpus of Historical English	Rissanen	1.5
7	1987	The Macquarie Corpus	Peters	1
8	1985	The Kolhapur Corpus of Indian English	Shastri	1

（续表）

序号	时间	语料库名称	创建者	容词量（百万词）
9	1971	The American Heritage Intermediate Corpus	Carroll et al.	5
10	1985	BCET	Sinclair and Renouf	20
11	1986	JDEST	Yang & Huang	3.6
12	1991	The Longman/Lancaster English Language Corpus	Summers	30
13	1992 -	CSAE (The Corpus of Spoken American English)	Chafe et al.	1
14	1988 -	ICE (The International Corpus of English)	Greenbaum	/
15	1992 -	BNC (The British National Corpus)	Quirk	100
16	1987	The Bellcore Lexical Research Corpora	Walker	250
17	1996	Bank of English	COBUILD	320

表 2.5 大型电子文档的发展(参见 Biber 1992)

序号	时间	名称	简介
1	1989	ACL/DCI (The Association for Computational Linguistics Data Collection Initiative)	"计算语言学协会数据收集行动旨在收集机读电子文本,以支持科学和人文研究"(Biber 1992).
2	1992 -	ACL/ECI (The European Corpus Initiative)	9300万词的多语种大型电子文档,其中部分语料通过光盘发行。
3	?	CLS (The Cambridge Language Survey)	剑桥语言通览为国际性多语种大型电子文档,主要用于语义分析和索引。
4	1992	LDC (DARPA-funded Linguistics Data Consortium)	语言学数据联盟收集大量的原始文本,以及已经进行句法和语义赋码的文本与数千小时的话音转写材料。
5	1979	American New Stories	25万词的美国英语书面语材料。
6	1988	The Nijmegen TOSCA Corpus	包含75部著作共150万词的英国英语电子文档。
7	1987	The Melbourne-Surrey Corpus	10万词的澳大利亚报刊英语电子文档。
8	1984	The Corpus of Canadian-English Writing	300万词的加拿大英语电子文档。
9	?	The Warwick Corpus	由J. M. Gill创建的250万词的英国英语书面语电子文档。
10	1980	The Cornell Corpus	由1151段语篇共160万词的英国英语和美国英语书面语和口语电子文档。
11	1980 -	NEXIS, LEXIS, and MEDIS	NEXIS主要包含报刊、时事通讯等语篇; LEXIS主要包含法律文书; MEDIS主要包含医学文献。

可以看出，在20世纪末，语料库发展进入了一个前所未有的阶段，“计算机语料库研究者们突然发现身处在一个不断扩大的世界”，“这种发展应使那些语料库的先驱者们感到欣慰。他们就像是从一辆驴车突然坐到了游行队伍中的一辆花车上”（Leech 1991:80）。

5 语料库发展目前存在的问题

从90年代至今，语料库发展可谓突飞猛进，基于语料库的各种研究也高速增长。根据乔恩森统计，基于语料库的研究在1965年以前只有10项，从1965年到80年代初15年间有130项，而从1980年到1985年5年间就有160项，超过了前15年的研究数量总和。1985至1991年6年间的研究数量又比前5年增长了一倍，达到320项（Johansson 1991，引自McEnery 1997）。但是，如今的语料库发展中仍存在一些问题，综述如下：

1）书面语语料库与口语体语料库的发展不平衡。一方面，大量在线电子语料以及通过各种电子媒介发行的电子文本为语料库的建设提供了无尽的资源，使书面语语料库容量以百倍甚至千倍的速度增长，同时也使各种专用语料库建设变得轻而易举。另一方面，口语语料库发展却不尽人意。原因之一是自然语言的日常运用难以收集，通过录音、摄像等手段进行采集存在许多法律和技术上的困难。其二是录音材料需要人工转写（transcription），耗费大量人力和物力。再则，口语语料的标注和赋码缺乏统一的标准格式，对该种类语料的数据分析也缺乏专用工具软件。以上种种因素制约了口语语料库与书面语语料库的同步发展。这也使书面语语料库的大小必须得到适当控制，以求各种语料的平衡（Leech 1991:73－80）。

2）语料库的标注和赋码系统的统一性和适用性应适当平衡。

在对语料库进行自动词性赋码时，赋码系统应能准确区分和标注自然语言中大量的同形异义词，如 *bark*（树皮）与 *bark*（犬吠），*bank*（河岸）与 *bank*（银行）等。现行的赋码方案一般是基于传统语法中对词类的划分，而这种分类只是基于一种"一致认可"的假定。里奇指出，其实任何一种分类方案都可能囿于某一语言学流派之学说，不免顾此失彼。所以，理想的赋码方案的确立，应是基于语料库自身的统计信息。里奇甚至设想，由全世界的语言学专家坐在一起，制定一个标准的赋码方案（Leech 1991:80）。然而，由于大量专用语料库的不断涌现，对语料库的标注和赋码也应以满足不同研究目的和应用领域的需要为基础，这就意味着允许不同的语料库用户设计自己的赋码方案。所以，理想的标注和赋码系统应该是一个开放的系统，其格式应力求标准化和统一，而具体赋码方案则可由用户自己制订。

3）语料库资源共享仍困难重重。一个语料库建成后，其价值与被利用的程度成正比。然而，目前除了少数几个语料库能够提供在线索引外（如 COBUILD、BNC、TEC），大部分已建成的语料库只掌握在小群体的语料库研究者手中，大多数圈外人只闻其名，难见其形。此外，仅有的能够提供在线索引的几个语料库，大多数功能单一，如只提供搭配词和词语索引，几乎只起到一种演示作用。汤普森（Thompson）在引用英国工商部 80 年代末签发的一份文件中指出，无论采用何种方法，大量的经过标注的文本数据对口语和自然语言处理研究都是至关重要的。但就目前而言，英国还不具有在规模、质量和使用上都适合使用的语料库。经过 10 年发展，语料库这种尴尬的局面并没多少改变。目前，常见的语料库资源共享方式有：(1)通过 CD-ROM 发布。如 COBUILD 和 BROWN 语料库等；(2)把语料库放在网络服务器上，并把语料库引擎放在网页上，供用户登录使用，如 COBUILD，BNC，TEC 等；(3)提供一个独立的客户端程序，用户下载后安装在自己的 PC 机上，通过上网进入远程语料库进行查询，如 BNC 的 SARA，TEC 的 Tec-Concordance 等。把语料库索引结果直接转换成超文本格式（htm

或 html, Hyper-text Markup Language)是一个不错的尝试(如 Web Concordance)[1]。自动生成的网页索引可以在互联网上发布。不足之处是,每次处理的语料有限。在将来的发展中,我们需要更为方便的渠道来共享语料库资源,在线语料库所提供的索引功能也应更全面。

4) 语料库工具软件和文本分析软件的开发与语料库自身的发展不协调。语料库工具软件包括自动赋码和句法分析等。文本分析软件则指与语料库本体相对独立的索引软件。从 80 年代至今,语料库工具软件与文本分析软件的核心技术并没有像语料库那样产生突飞猛进式的发展。各种索引软件重复开发,功能单一,软件升级速度缓慢,且自由软件极少。其主要原因在于:(1)语料库软件开发不像其他计算机软件开发那样有巨大的商业利益驱动,几乎没有专业的软件开发人员。大多数开发者往往是精通计算机编程技术的语言学家,其软件难以跟上最先进的计算机技术发展。如使用 WINDOWS 操作环境的索引软件直到 1997 年才面世,而当时微软的 WINDOWS 操作系统从 3.x 到 WINDOWS 97 已更新换代了好几代。(2)除少数软件(如 TACT)是由多伦多大学多个统计学家和软件工程师组成的开发小组共同研制,大多数语料库工具软件和分析软件都是单兵作战,研发过程漫长,软件缺陷多。这主要是由于语料库软件开发缺少投入,难以吸引优秀的软件工程师。另外,软件发行渠道不畅通,用户难以获得,这也是阻碍其发展的主要因素之一。(3)文本分析软件功能单一,且对其分析过程和结果缺乏必要的说明和解释。例如,自动赋码技术如今已能获得 98% 以上的准确率,而如何充分利用赋码信息进行词性与搭配分析却难以见到有效的解决方案。(4)语料库对外语课堂教学无疑具有巨大潜力,但是对于如何充分利用语料库与词语索引,使之服务于外语教学和学习(如"数据驱动学习"DDL),缺乏必要的应用开发。随着 21 世纪的到来,网络多媒体辅助外语教学将普遍展开,语料库技术作为一种先进的学习工具和资源产品,势必广泛应用到整个教学过程中。所以,语料库在外语教学中的应用将成为外语教学中的主要

课题之一。

6 中国英语学习者语料库

学习者语料库作为一种专用语料库的发展也只是近十几年的事情。学习者语料库通过收集语言学习者各种书面语和口语的自然语料，试图建立一种学习者语言数据库。现代计算机学习者语料库(Computer Learner Corpus，简称 CLC，参见 Granger 1996)与以往的学习者错误语料库不同，其目的在于对学习者的语言特征和语言发展进行全面而系统的描述和对比分析，而不是仅限于对学习者错误进行分析。学习者语料库的优势在于能够提供有关学习者语言发展的全面信息。在外语学习和教学中，基于学习者语料库的分析能提供学习者的典型困难以及在某一具体方面的主要障碍的反馈信息。另外，通过不同类型学习者语言对比，可以发现语言学习者在某一发展阶段的共同特征和个体特征，从而促进外语教学在大纲设计、教材开发以及课堂教学诸环节更适应学习者的需求。

现代学习者语料库往往与学习者中间语(interlanguage)分析联系在一起。学习者语言不再被简单看成是一种“错误”，而是普遍存在于学习者中的一种规则系统(参见 Selinker 1972，Corder 1981)。格兰杰(S. Granger)把基于学习者语料库所进行的中间语分析称为“对比中间语分析”(Contrastive Interlanguage Analysis，简称 CIA)。她认为，这种分析能为外语教学提供极有价值的资源和信息(Granger 1996)。中国英语学习者语料库从 1996 年开始筹建，于 1999 年年底完成。该语料库由专业英语、大学英语以及中学英语三种类型的学习者的书面语料构成，同时提供生语料库和赋码语料库两种版本。本节通过中国英语学习者语料库建设，说明语料库建设的基本过程和方法。

值得注意的是，学习者语料库中的语言运用材料的“自然性”

(naturalness)不同于一般意义上的语料库。无论其来自自由作文、指导作文、试卷作文还是其他方面,其语料都不免带有“诱导”(elicitation)特性,不能算纯粹的“自然”语料。所以,格兰杰认为,学习者语料库数据只能尽量追求自然(1999:2)。

6.1　CLC:新兴的事业:学习者语料库的发展

CLC 是 Computer Learner Corpus 的简称。现代学习者语料库是通过收集语言学习者各种书面语和口语的自然语料而建立起来的一种学习者语言数据库。与以往的学习者错误语料库不同,其目的在于对学习者的语言特征和语言发展进行全面而系统的描述和对比分析,而不是仅限于对学习者错误进行分析。随着计算机处理速度的飞速增长以及存储能力的扩大,语料库建设呈现多元化发展和增殖趋势。在 90 年代,各种专门语料库纷纷建成,如用于文学作品分析的各种作家语料库,用于历史语言学分析的历史语言语料库,以及用于学习者语用分析的学习者语料库。语料库的种类不再仅限于英语,而发展到各个语种。语料库多元化的发展既包括研究方法的多元化,如历时性语料库和共时性语料库,也包括语料语体上的多样化,如书面语和口语语料库、通用型语料库和专门语料库,以及英语的各种变体语料库,如英国英语、美国英语、澳大利亚英语、印度英语、南非英语等。在这种多元化发展中,针对外语学习者而建立的学习者语料库(Learner Corpora)可谓异军突起,成为当今语料库建设中一股新的力量。学习者语料库的创建和研究只是近几年的事情。最早的学习者语料库是 80 年代末建立起来的朗文学习者语料库(Longman Learners' Corpus)。90 年代中期,在比利时卢弗安大学以格兰杰为首建成了国际学习者英语语料库(ICLE)。该库是一个广泛的国际合作项目,容词量为 100 万词,所搜集的学习者语料来自 14 种不同的母语背景(包括法、德、荷兰、西班牙、瑞典、芬兰、波兰、捷克、保加利亚、俄、意大利、希伯来、日和汉语)。另外,香港科技大学的学习者语料库(HKUST Learner Cor-

pus)搜集了以汉语为母语的学习者语言材料,容量为360万词。学习者语料库的发展,使得基于学习者语料库的研究异彩纷呈,方兴未艾。1996年8月在芬兰举行的"第十一届世界应用语言学大会:开发计算机学习者语料库"上,与会者从不同角度对学习者语言进行了初步探讨。如阿茨(J. Aarts)对多义动词 *find* 和 *want* 的对比研究,阿尔滕伯格(B. Altenberg)对瑞典英语学习者议论文写作中各种词法、句法和话语特征的滥用或少用的研究,卡斯苏布斯基(P. Kaszubski)对波兰英语学习者词汇的重复和华丽语句的运用所进行的对比研究,洛伦茨(Lorenz)对德国英语学习者英语写作中词汇搭配能力、非词汇化以及信息结构的研究,密尔顿(Milton)则讨论了机助语言学习设计问题;林伯姆(H. Ringbom)对比分析了学习者语料库中数量限定词、核心形容词和动词、衔接词以及动词短语的频率。

到目前为止,国际上建成的其他英语学习者语料库有日本的朝雄幸次郎(Kojiro Asao)负责建成的 *Corpus of English by Japanese Learners*、宫昭二(Shogo Miura)的 *Japanese Junior High School Students' Japanese/English Translation Corpus*、匈牙利潘诺尼斯大学的 *Hungarian EFL Learner Corpus*,以及 *Cambridge Learners' Corpus*[12]。

1999年,在中国建成的容词量为100万词的中国学习者英语语料库(CLEC),广泛搜集了专业英语、大学英语、以及中学英语学习者的各种书面语资料。该项目由广东外语外贸大学桂诗春教授和上海交通大学杨惠中教授主持,由国内十几所院校合作完成,目前已完成人工错误赋码和校对工作。

6.2 中国学习者英语语料库建设的目的和意义

就我国外语教学而言,建立学习者英语专用语料库,并以语料库为基础开展学习者语言分析,无论是对我国21世纪外语教学,还是对我国语言学与应用语言学研究,都具有重大的意义:(1)对我国大学英语教

学和学习而言，英语写作和口语属于语言表达能力。根据统计研究，学习者的写作能力与他们的整体英语能力密切相关。换言之，根据学习者的写作水平往往可以衡量其总体英语水平。因此，从分析学习者所写英语入手，观察他们中间语(interlanguage)的发展(Selinker 1972)，进而分析母语在英语学习中的正迁移(positive transfer)和负迁移(negative transfer)现象，为我国外语教学、尤其是写作教学，提供大量的可靠数据并产生积极的后效。(2)语料库的建立和开发将对各种标准化英语测试和作文评分提供坚实的依据，为题项的命题和题项测评提供客观数据，这将有助于提高题项的结构效度和评分信度。(3)对我国外语教学研究而言，语料库的研究开发将有助于推动量化研究的发展，深化外语教学研究中"数据驱动"这一原则，使教师和研究者掌握这一有力工具，结合自身经验和语言直觉，真正体现外语教学的科学性、艺术性和技术性。(4)与国际外语教育研究接轨并达到同步发展的水平。真正意义上的与国际同步发展，是应用国际最新成果和先进技术手段，结合我国外语教学实际，开展面向我国外语学习者的研究，而这些只能由我国的外语工作者来实现。在即将到来的21世纪，我国将面临更强的国际竞争，国内外人才市场对人才的外语水平提出更高更强的要求，我国的外语教学无疑必须适应并满足新的需求。

中国学习者英语语料库(CLEC)初始容量为100万词，在设计时就规定了严格的设计标准，对可能影响外语学习的各种变量通过标识进行控制，如学生类型、性别、学习积累年限、完成方式等，在大学英语子语料库中还赋加了诸如作文分数、水平等级等信息。与其他学习者语料库不同的是，该语料库还设计了一套错误赋码体系，对所有语料进行机助人工赋码，使其反映的信息更集中和更具有针对性。

在创建自己的语料库前，首先应就语料库的用途和创建过程确定一系列基本原则，包括语料库的大小、类型、语料覆盖范围、抽样方法、语料处理手段、可用索引软件等。另外，还要决定是否开发专用的索引软件。如果不需要开发专用索引软件，就要预先考虑好主要使用何种现有的索引软件，并根据该软件的特性和功能确定语料的标注格式和赋码格式。语料库一般由库本体(语料文档)和语料

库引擎组成。语料库本体由根据语料库用途和建设原则进行随机抽样的大量电子文档组成,其存储格式一般为纯文本。语料库引擎即对语料进行处理和索引的专用软件。传统的语料库往往与所用引擎集成在一起。这种索引软件只能识别该语料库特定格式的标注和赋码,语料库本体与索引软件构成了一个封闭系统,对外来语料的处理只有通过电子文本特定的标注才能进行。现在,不少索引软件独立于特定的语料库,这使得语料库建设者只要专心于语料的收集与整理,而不必进行索引软件的重复开发。另外,随着互联网的发展,大量在线资源和电子文档为语料的收集提供了极大的方便,加之不少功能强大的索引软件已可以通过网上免费下载,使得原来耗时费力的语料库建设变得轻松易为。如今,一个外语教师或语言研究者为满足自己特殊目的建立一个专用语料库,已不是一件很困难的事。以下仅以中国英语学习者语料库为例,简略说明语料库开发的过程以及需要注意的问题。

6.3 CLEC 语料库的设计原则

由广东外语外贸大学桂诗春教授和上海交通大学杨惠中教授牵头开发的中国英语学习者语料库 CLEC 开始建于 1997 年 4 月。当时,我们建立该语料库的主要目的有以下两个方面:1)通过分析中国英语学习者写作中典型错误及其与学习者中间语发展的内在关系,为中国外语教学,尤其是英语写作教学,提供积极反馈;2)对学习者语料库与英语本族语语料库进行对比分析。在创建初期,首先确定该语料库在性质上属于专用语料库。语料库的初始容量为 100 万词。语料构成包括专业英语学习者作文、大学英语学习者作文、中学英语学习者作文以及其他英语学习者作文。所以,CLEC 语料库建成后,将由专业英语、大学英语和中学英语三个子语料库构成。语料抽样按照比例抽样与分层抽样相结合的方法进行,即语料类型根据不同的英语学习者按比例抽样(见下图),而对同一学习者群体的英语作文按不同层次进行抽样(见下图)。

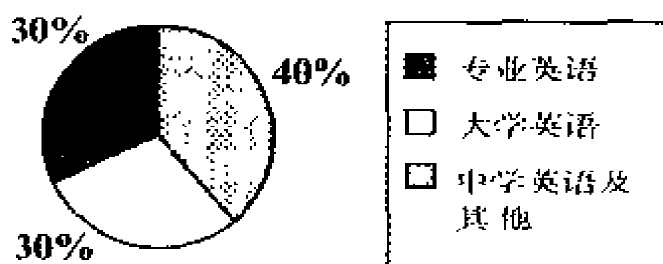


图 2.5　CLEC 比例抽样示意图

针对大学英语学习者语料库建设，杨惠中教授提出了以下几条原则：1）首先建立原型，即先从大学英语作文中抽取少量作文，并按分数段进行分组，目的在于对错误类型和典型特征进行归类划分，并以此作为以后制订赋码方案的依据。2）对错误类型的划分可按不同层次进行，如词法、句法、搭配、语义以及语篇等。语料库建成后，可就以上错误类型的分布与作文的得分信息进行相关分析。3）分析结果可作为大学英语考试作文部分给分的客观依据。4）建成后的语料库应能实现网上资源共享。5）如有可能，亦可对作文中地道的表达和特征进行赋码。根据这些原则，语料库建设的初步工作包括对大学英语学习者作文进行小样本抽样（30 份）以确定错误特征、根据抽样分析制订初步的赋码方案、作文收集、键入计算机、校对、赋码以及赋码校对和对比。

正如格兰杰（1996）所说，只有设计优良的语料库才会产生好的结果。在语料库设计阶段，不仅要确立基本设计原则，还要考虑到为语料库的使用者能提供哪些可供查询的信息。在学习者语料库中，有关语料以及作者本身的重要信息必须在语料标注中体现出来，如作文类型（包括自由作文、试卷作文、日记、指导作文等）、学生类型（包括专业英语学习者、大学英语学习者、中学英语学习者等）、个人信息（包括学习年限、性别等）以及语料信息（包括题目、是否使用词典等）。在大学英语学习者子语料库中，语料标注还包括了学习者分级信息（四级或六级）以及评价信息（作文得分）。在语料库分析中，不同得分的作文可以解释为学习者外语水平的差异。杨惠中教授认为，此类信息非常重要，因为它可以使我们对比不同水平

的学习者在英语写作中的典型差异，并依此观察学习者中间语的各个发展阶段。他还预言，各种错误类型在不同水平的学习者作文中存在差异，语言形式错误的分布可能较集中在低水平作文中，而较高水平的作文中出现更多的则是语用错误。这一预言在以后的统计分析中得到了证实。

需要注意的是，在分配各种语料类型的比例时，应充分保证语料的代表性。目前的大学英语子语料库只包括了试卷作文和自由作文两种类型。这两种类型的语料各有优缺点。一方面，试卷作文能够真实反映学习者目前的写作水平，考场的压力使他们无暇他顾，只能发挥平时的能力。另一方面，学习者在考场中的压力和焦虑感使他们的作文并非常态的语言运用。其数据的“自然性”较弱。与此相比，自由作文似乎更符合学习者语言运用的常态，所得数据更自然些。但在语料校对过程中，我们发现有些自由作文含有从现成文本中照搬的句子和片段，这些句子很难反映学习者真实的语言水平。所以，在校对过程中，要把那些含有大量出处可疑的句子或片段的作文滤除掉。另外一个问题是，由于所有语料都是手写文本，必须通过键盘输入到计算机，所以初步语料含有大量的输入性错误，如增加或遗漏字母或词汇、标点识别错误以及空格键或非法字符的插入等。这些错误如不及时纠正，将会严重影响语料处理的可靠性。

6.4 语料抽样

以大学英语学习者子语料库为例，首先抽取小样本试卷作文以确立错误原型，并以此为基础确定抽样的范围。在小样本作文抽取时，等量抽取从 5 分到 15 分(满分)作文作为分析基础。我们发现，6 分以下的作文中完整句子很少，达不到抽样要求。所以，我们决定只抽取从 6 分到 15 分的大学英语四、六级试卷作文，并以 1996 年 1 月至 1997 年 1 月三次考试的试卷作文为时间范围。采取随机方法对每次考试等量抽取 300 篇作文。实际结果是，四、六级作文平均各抽取了 1000 篇。后来又追加两种作文共 500 篇，加上收集的 1000 篇自由

作文，构成了容量为50万词的大学英语子语料库。这样做的目的是尽量保证语料对不同层次和水平的学习者总体的代表性。为了说明问题，就试卷作文部分各个分数段的作文抽样量作图，以观察其分布情况（见图2.6）。从图中可以看出，各个分数段的取样作文分布基本呈正态，峰态集中在8分和9分段。整个直方图稍呈负偏态，这是由于取样分段不是从最低分开始抽取，而是从6分开始抽取。这种分布基本符合我们抽样时的期望，即大多数作文应集中在8分到10分之间，以代表大学英语学习者的平均英语写作水平。

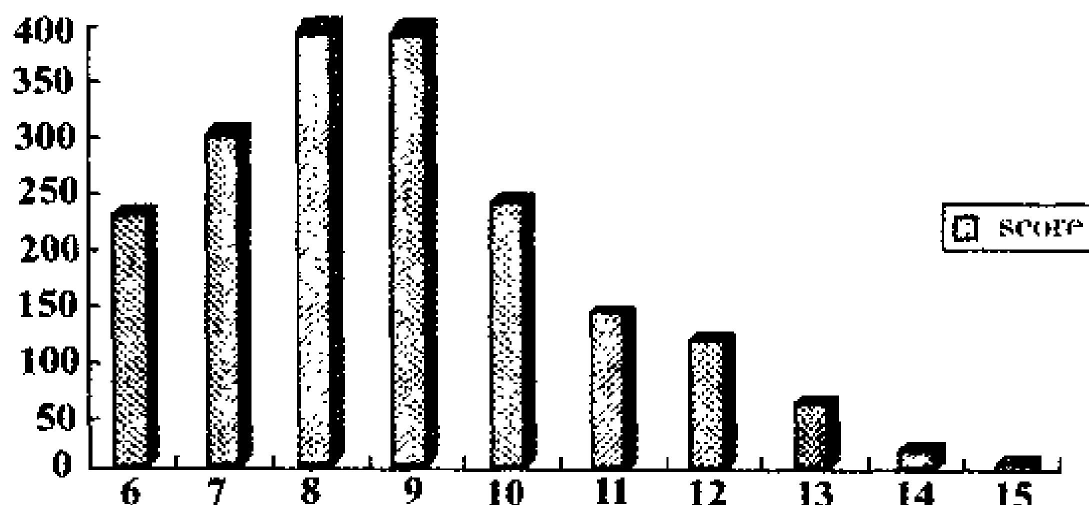


图2.6　取样作文在各个分数段的分布

6.5　标注与赋码

根据桂诗春教授提供的方案，对作文的编码标注包括以下8种主要信息，标注编码采用COCOA标准，一律放在尖括号“〈〉”内，并置于每篇作文的第一行，以便软件识别：[13]

ST（学生类型）：下设7个参数，分别代表初中学生和高中学生、非专业英语四级和六级、专业英语一、二年级和三、四年级以及研究生，以从1到7的不同数值代表。如<ST 1>表示该篇作文的学生类型为初中生。

SEX（性别）：分别由1和2两个数值代表男性和女性。

Y（累计学习年限）：用数值代表。
AGE（自然年龄）：用数值代表。
WAY（作文完成方式）：1 表示试卷作文；2 表示课堂作业；3 表示课外作业。
DIC（是否使用词典）：1 表示是；2 表示否；？表示不确定。
TYP（作文类型）：1 表示论说文；2 表示叙述文；3 表示应用文；4 表示说明文。
在大学英语子语料库中，增加了以下编码标注：
SCH（所在学校）
SCORE（作文得分）：从 6 分到 15 分。
TITLE（作文标题）：原作文题目。
BAND（大学英语四、六级试卷作文编码）：4 表示四级；6 表示六级。
例如，根据以上方案，下列编码标注 <BAND 4> <ST 3> <SEX ? > <Y ? > <AGE ? > <WAY 1> <DIC 2> <TYP 1> <SCH ? > <SCORE 10> <TITLE Haste Makes Waste> 表示该篇作文属于大学英语学习者一、二年级学生的四级试卷作文，未使用词典，作文得分为 10 分，标题是"Haste makes waste"，其他不确切信息用"?"表示。

在制定错误赋码方案前，我们首先通过抽样分析，确立典型的错误特征，并依此对不同特征进行分类和编码。对赋码方案的制订，杨惠中教授提出了以下几个重要原则：1）每个编码应能表达最大信息量，并与所表达的错误类型直接联系起来。2）赋码方案应便于操作和修改，即该方案应为一开放系统，便于将来的修订和更新，并最大限度地满足将来的研究需求。3）编码的复杂度应与赋码效率平衡起来，即编码应表意明确、结构层次分明，不增加赋码者的记忆负担。4）应保证单个赋码者前后的一致性和多个赋码者之间的一致性。这些原则为以后不同赋码方案的合并和最后方案的确定发挥了指导性作用。最后，在桂诗春教授的主持下，经过数次讨论，并经其确认，制定了最后的赋码方案。该方案由 61 个错误编码构成。整个赋码方案为等级结构，分为句子、搭配、用词、动词、名词、

形容词、代词、介词、连词、词形等主要错误类型，每一主要类型下再分若干个子类型。如搭配错误为一主要错误类型，下分名(词)名(词)搭配、名动搭配、动名搭配、形名搭配、动副搭配、形副搭配等子类型。词形错误下分拼写、造词与大写三个子类型。约定赋码格式为所有赋码放在中括号[ ]内。赋码编码由三个主要部分组成：主要错误类型、子类型、该错误的覆盖范围(见示意图)。

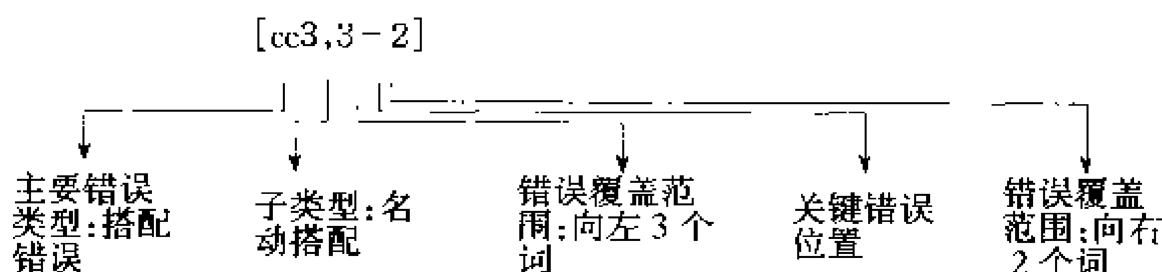


图 2.7　错误码释义

以下索引行显示在作文中插入赋码的位置：

1	uit different job so as to adjust competition	[cc3.1－]in the future society. [sn8,s-] It is tr
2	atever jobs they take and make responsibility	[cc3.1－3] for other perpons, [fml, －] what's more,
3	ssibly have great development or gain advances	[cc3.1－] in their field so long as they work hard o
4	wd7, s-]. For instance, when we do examination	[cc3.1－], because we feel like doing it at a quick s
5	tance, someone is haste to get [wd3, －1] train	[cc3.1－]. The time is rather limited, so he is upset
6	they hope to continuously renew their lives.	[cc3.2－] In my opion, [fml, －] I think job [np6, －
7	an be used and you can advance your business.	[cc3.2－] That's do [vpl, 1－] you good and that's i
8	] new surroundings and to study the knowledge	[cc3,2－] about his new job. In fact, he can make a
9	d the other kind. Because that can improve you	[cc3,2－] quickly! [sn2, s－] (Band 6)(SEX ?)(Y
10	he present job can't appear their capabilities	[cc3,2] or can't often [wd4, 2－2] high salery [fml
11	yment. Some people likes to transforms his job	[cc3,2－]. Because he can get higher salary or he ca
12	8, -s]. These people don't transform his work	[cc3,2], they can avoid unemplcyment. Some people
13	they have learned did not measure their job	[cc3,2－], [sn1, －s]. In my opinion, taking a job is t
14	,-1] pratices. And when you contact more work	[cc3,2－], you can find which [pr5, 3－2] you like.
15	te. Because we are too late to catch the class	[cc3,2－] on time, so we want to how to quickly to sa
16	other example, when you learn to play computer	[cc3,1－], it is certainly that at first you don't kno

图 2.8　错误码在文档中的插入位置

在以上索引行中，放在中括号中的编码便是错误码。错误码的位置一般在出错的关键词后。如在第一行中，错误码[cc3,1 -]放在名词 *competition* 后，[cc3]表示该错误属于搭配错误中的动名搭配错误，后跟的数值"1"表示该词左边第一个词与该错误相关联，而符号" - "表示该错误关键词出现的具体位置。通过专用软件处理，可以提取出 *adjust competition* 这一搭配错误。

由于错误码数量大，人工赋码很难记忆每一个错误码所代表的确切意义，且每次赋码靠人工键入可能会产生大量键入错误，所以有必要开发一个机助人工赋码工具，以减轻赋码者繁重的劳动。该工具必须能够免除赋码者对每一具体编码的记忆，以使他能够专注于作文中错误的识别和确认。另外，该工具应便于操作和移植，并适用于通用的操作平台。基于这一目的，建库者利用常用的 MS WORD 制作了机助赋码工具(见下图 2.9)，该工具可存入 MS WORD 的通用模板文件，轻松实现用户间的共享[14]。

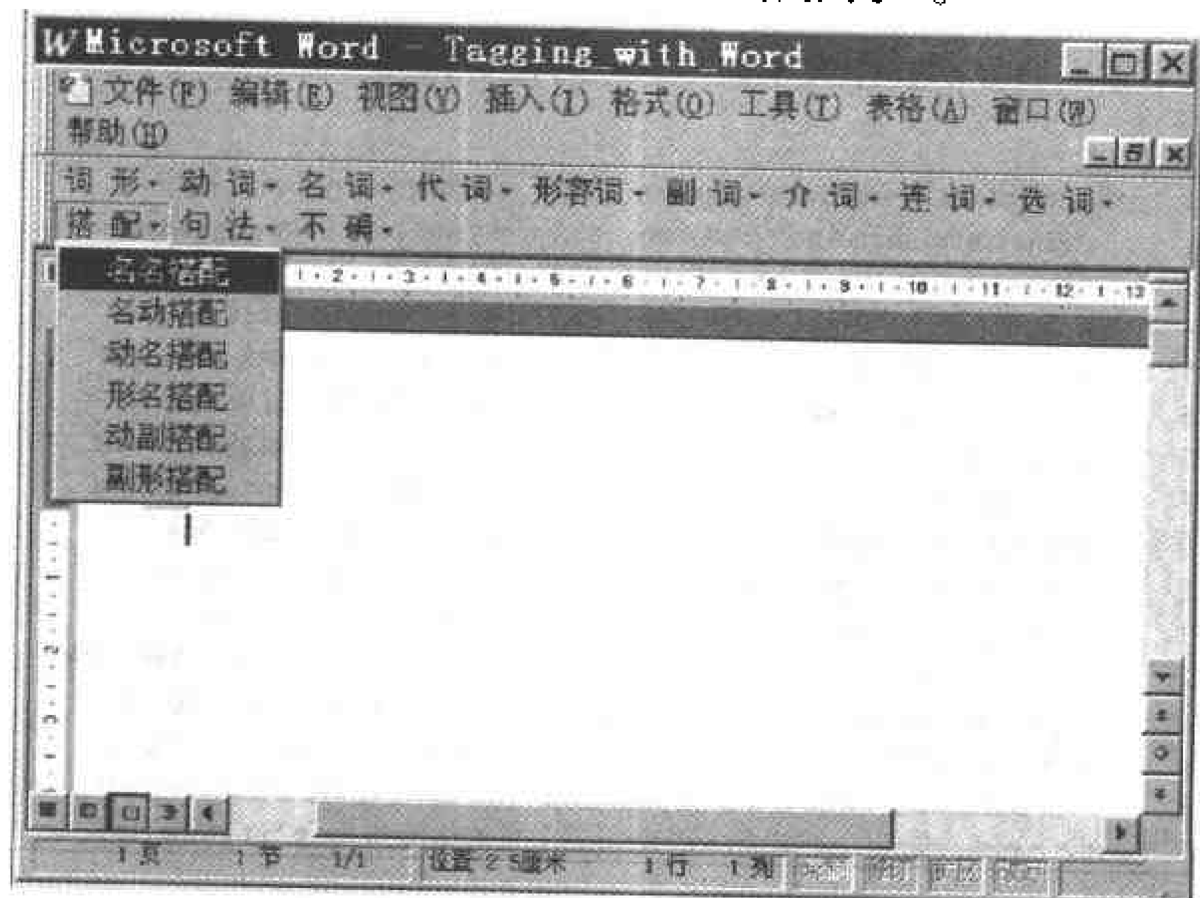


图 2.9 机助赋码工具 Tagger-with-Word

利用该工具，赋码者在赋码时，只需点击所选类型，代表该错误类型的编码即自动插入文本，无须重复查询编码（该工具可从李文中主页免费下载：*http://grwy.online.ha.cn/liwenzhong*）。

6.6　赋码校对与一致性检验

为保证赋码的可靠性，有必要在初步赋码完成后，重新校对所有赋码，并利用统计手段检验赋码的一致性。这里所说的一致性有两个方面的含义，一是同一赋码者前后赋码的一致性，另一是不同赋码者之间的一致性。由于同一赋码者在不同时间或不同赋码者对某种错误类型的主观判断与分类存在差异，往往造成赋码的偏差。如果这种偏差过大，会严重影响赋码的效度和可靠性。为避免此类问题发生，赋码者要经常对不同时间的赋码进行比较和矫正。不同的赋码者应尽量多讨论，不断相互对比，取得一致意见。赋码完成后，利用统计手段对各自的赋码总体情况进行对比分析，以保证赋码的效度和信度。例如，大学英语学习者子语料库主要由濮建忠和李文中两人完成，两人所负责赋码的语料数量和类型基本一致，具有很强的可比性。他们采取的检验方法是，统计各自所赋各个错误类型码的总量，然后就其分布进行对比分析，如下图：

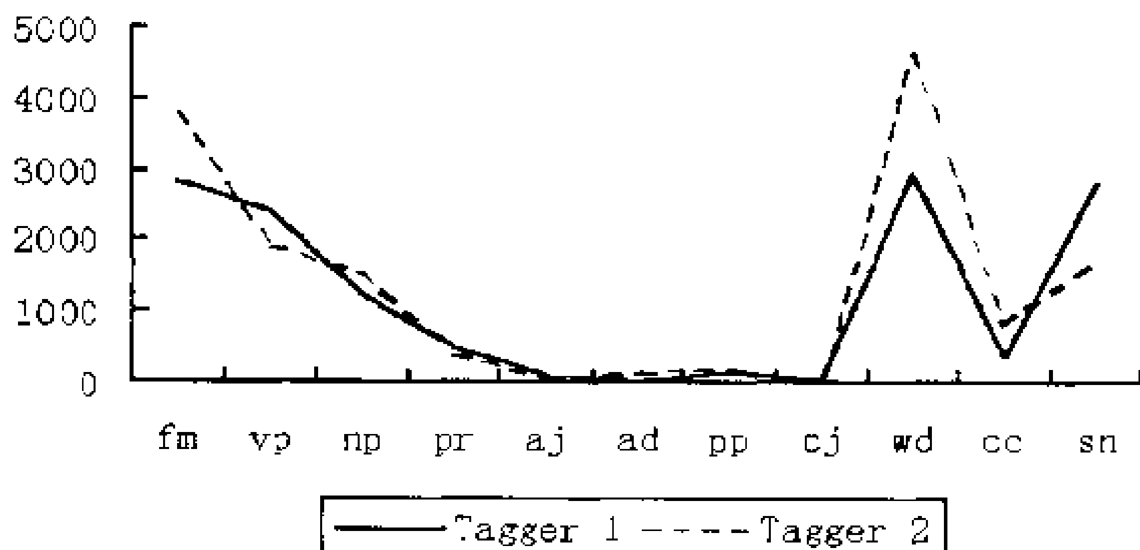


图 2.10　两个赋码者结果对比

从图中可以看出，他们两人所赋各个错误类型码的结果具有很高的一致性。图中两条曲线基本吻合，只存在少许偏差。可以说，该语料库的赋码是可靠的。

7 学习者语料库在外语教学中的应用

学习者语料库的意义在于：1)在语言学习环境方面明确区分外语学习和第二语言学习，从而观察和描述不同的母语背景与目的语接触程度差异对语言学习的影响。2)对"学习者语言"进行全面而系统的调查和描述，并通过与本族语语料库对比，确认学习者的主要困难，以期对外语学习和教学产生积极的反馈效应。3)对于语言习得研究而言，对大量的学习者语言运用材料进行量化分析，能深化人们对语言学习机制的理解，从而对语言本身的理解提供依据。事实上，语言习得研究的主要数据依据来自三个方面：1)学习者的语言运用；2)研究者为某一研究目标从研究对象引出的信息；3)学习者通过内省而提供的信息(Ellis 1986)。传统的语言习得研究由于受研究手段和人工处理信息能力的限制，第一种信息的获得和数量难以满足研究者的需求，所以主要依赖后两种信息。如今，语料库技术的发展为解决以上问题提供了有效的途径。与传统的对比研究(CA)不同，利用学习者语言与本族语进行对比，也可以在不同的学习者语言之间进行对比(如不同母语背景的学习者在学习困难上的差异)，所得到的信息更加可靠。基于学习者语料库的分析也不同于传统的错误分析(EA)，研究者不仅可以分析学习者的语言形式错误和语用错误，还能通过对比分析进一步观察学习者使用规避策略(AVOIDANCE)的情况。格兰杰(1996)把这种对比分析称为中间语对比分析(CIA，Contrastive Interlanguage Analysis)，并认为通过这种对比不仅能发现学习者语言中不合乎本族语的特征，还能发现某些特征在学习者语言中的滥用或少用。对于外语教师而言，学习者语料库将能提供直接方便的帮助。这种帮助包括实时错误分析、机

助教学、针对性个人练习、学习评价等。本族语(目的语)语料库为外语教学提供更为真实可靠的语言材料,在教学大纲制订、教材开发、词典编纂等方面将发挥愈来愈重要的作用。但在此过程中,仅依赖英语本族语语料库是不够的,因为英语本族语语料库只能为我们提供英语语言典型结构和用法的可靠信息,却难以说明这些结构和用法对于学习者的难度。所以,外语教学的改进,不仅有赖于真实的目的语材料,还有赖于真实可靠的学习者的信息。前者能提供典型的英语运用信息,而后者能够告诉我们在英语学习中,哪些困难是普遍性的,哪些困难只存在于某一学习群体(Granger 1996)。兰德尔(M. Rundell 1996)批评那些英语教学材料的编写者依然依赖自己的直觉和教学经验,并根据现成的所谓"不证自明"的语言事实编写教材。他断言,正如一部教学词典如果没有可信的语料库依据就很难在市场上立足一样,人们愈来愈期望英语教学材料对英语的描述也应该具有实验性依据。大量基于语料库的研究结果表明,不少词汇意义和用法在真实的语言材料中的表现与传统语法和词典的定义和描述并不一致。语料库的意义在于能为我们提供大量真实语言材料,这些自然的语言结构和用法在语言现实中是完美的、可以接受的,但对于那些采用内省法的研究者而言可能是不合乎语法的,而他们从自己直觉出发造出的句子却难以在语料库中得到印证(Edwards 1993)。

另外,语言学习的过程不再被视为习惯的形成,学习者也不再是被动地对刺激作出反应。在当代交际语言教学运动中,外语学习更加注重学习者的个人需求。学习者通过语言接触自己发现规则,作出假设,并在语言运用中不断检验和修正自己的假设(Corder 1984)。语言学习也不再被视为教师—学生之间的单纯的知识传授。随着学习者语料库和各种专门语料库的发展,外语学习者和教师将得到大量的语料资源和在线帮助,学习者的外语接触和语言输入将远远突破以往的限制,困扰语言学习的"真实材料"问题和真实交际问题将得到有效解决,使外语学习更富于交互性和个人化。值得一提的是,语料库在外语教学中的应用不仅不会削弱教师的作用,相反,它为教师提供了更为广阔的创造空间。方便快捷的自动处理使得教师有更多的精力

和时间致力于教法的改进和设计更富创造性的教学活动。语料库本身并不是目的,而是一种更好的手段或工具。

综上所述,外语本族语语料库与外语学习者语料库结合在一起,在外语教学中具有广泛的应用前景。首先,语料库词语索引为现代"数据驱动学习"提供真实语料资源和工具。利用语料库词语索引可编制辅助语言学习软件,使学习者通过实际观察语境和实时练习,掌握某一语法结构或词语的用法,如约翰斯制作的 DDL(数据驱动)学习软件 Xcontext 可作为这方面的一个典型实例。该软件属于计算机辅助学习课件,基于语料库词语索引开发。打开程序后,学习者可在主菜单中选择要练习的词类,并可通过菜单提示获取帮助信息。选择某一词类后,进入下一屏幕,显示在该词类下可供观察练习的词汇。选择"Research"键可得到有关该词汇的问题提示。学习者通过观察所提供的词语索引行,分析研究某一词在语法和搭配中的典型用法,并通过对该词的实时练习掌握它。该软件还提供各种小测试,让学习者通过大量实例检验自己对某一词汇的掌握程度。第二,利用语料库,可以进行课堂词语索引,通过学习者参与,可以实时检索有关词汇和语法信息,并对学习者某一典型困难进行分析。第三,利用相关软件,可以把语料库全部词语索引结果转换为网页发布,供教师和学生在线检索和学习(见图 2.11)。第四,利用语料库词语索引制作网络多媒体课件。

在网络多媒体课件制作中,也可以结合相关软件,编制实时练习,使网页更富于交互性。下图(图 2.12a)是利用 JDEST 语料库索引制作的交互式填空练习。该练习使用超文本和动态超文本技术制作动态网页,使学生可以通过在线课堂进行英语学习。这种动态网页可以通过鼠标跟随技术和框架页提供实时提示和即时反馈信息(图 2.12b)。在外语教学过程中,其基本思想是教师既是组织者,又是教学工具的开发者。其主要特征是外语教师在应用中能够充分发挥创造力,集中精力利用可用资源,通过基于语料库的的课件开发向学习者尽量提供语言的真实运用的典型实例,使他们在自我探索中真正掌握目的语在词语、用法以及搭配语境诸方面



* 本网页使用 R.J.C. Watt(1999)的 Concordance 1.1.3 制作。

图 2.11　大学英语学习者语料库在线索引

的系统知识，并在实际运用中获得交际能力。另外，对广大外语教师而言，实现以上诸环节的技术只能是从属性的，并且容易学习掌握。这样才能使作为开发者和应用者的教师把主要精力放在内容的选取和呈现上，而不是迷失在复杂技术操作中。

语料库与学习者语料库在外语教学中的应用是一个尚待进一步开拓的领域，随着这一应用过程技术环节的简化和语料库资源的普及，语料库最终将成为普通外语教师手中的常用工具。

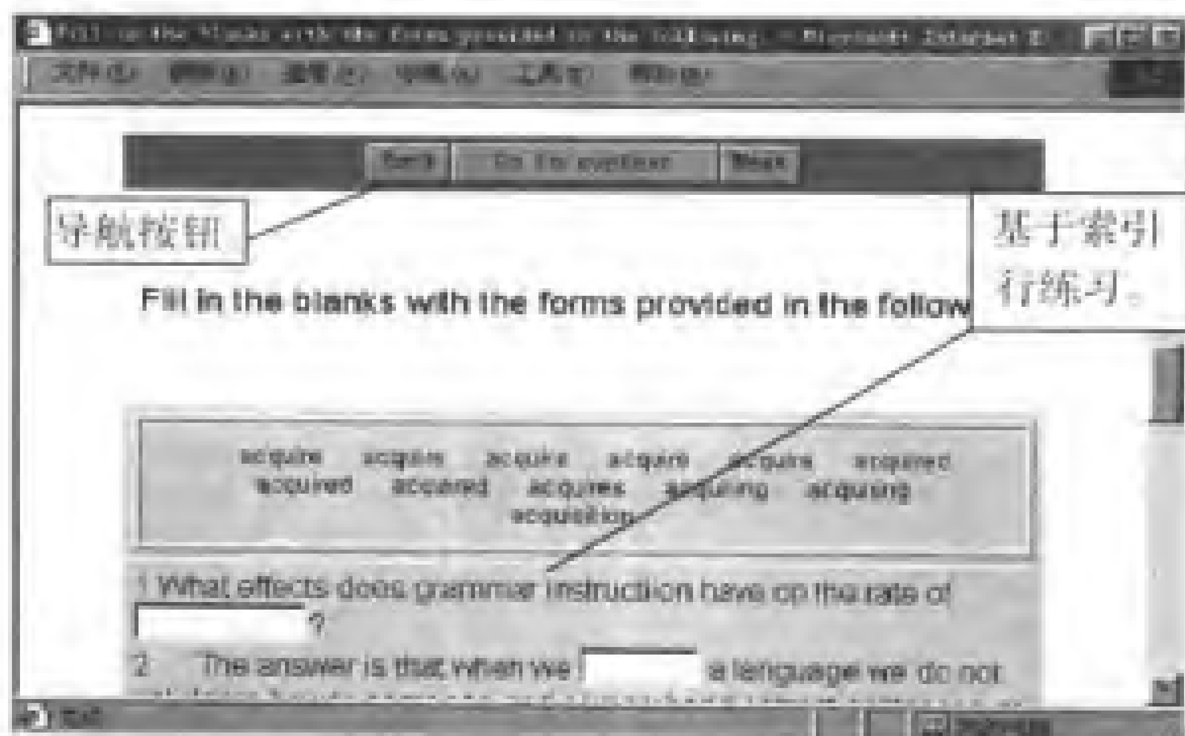


图 2.12a 基于语料库索引的实时在线练习

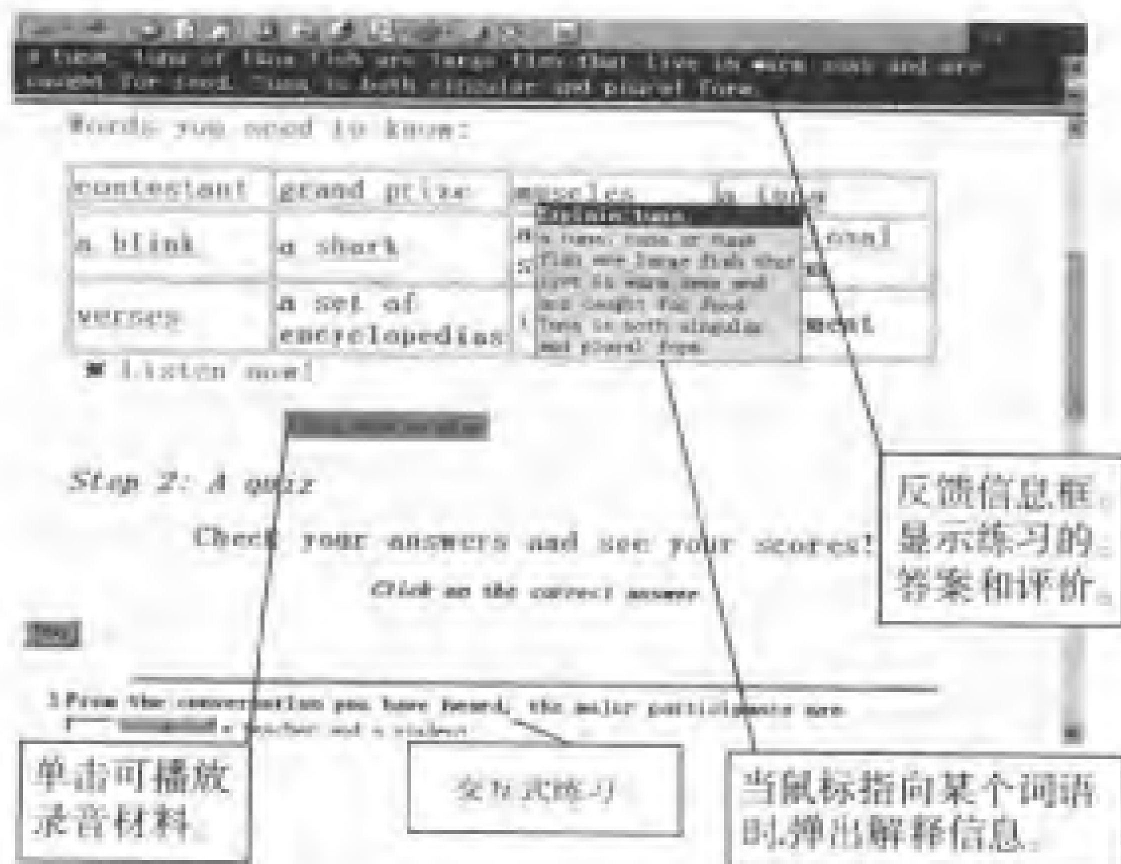


图 2.12b 基于语料库的在线学习课件

8 问题与结论

目前,国内一些高校和研究机构正在建设或准备建设用于各种目的的语料库。根据我们建设语料库的经验,在语料库建设过程中,以下几个方面的问题需要在语料库建库真正开始之前就认真考虑,作出决策[15]:

1）选择语料的基本原则。制订选择语料的基本原则应力求详实可行。需要进行的决策包括语料信道(如书面语还是口语,或二者混合)、语料的格式(即语料是公开发行的还是未出版的,是否需要获得作者或出版者的许可)、潜在用户(即语料库建成后,谁可能使用该语料库)、语料获得渠道和方式(如是否需要手工输入或扫描;电子文本是否需要格式转换等)、语料性质(即语料属于事实性信息还是虚构性信息或者是介于二者之间)、建库目标(即该库是用于论证、描述还是提供信息)、主题(即语料的主题范畴)。对以上问题的决策将影响到所建语料库的基本构成和应用趋向。

2）文本编码格式。建议把所有电子文本统一储存为纯文本格式或带分行符的 DOS 文本。这样做的好处是建成的语料库对各种索引软件适用性强,语料处理准确性高。某些软件不能识别其他编码格式的文本,对一些特殊格式标识符号在读取中会出现乱码,从而影响处理结果的精度。另外一种比较好的做法是对同一电子文本做几种不同编码格式的备份,然后分别比较某一索引软件的处理结果。

3）语料库准备。语料库准备步骤包括语料清理、语料分割、词性赋码、语义赋码等。语料清理是指对语料的分类、筛选以及各种语料的分配比例的确定和检查。语料分割即把语料分别存为较小的文档。分割的标准既可按文档的大小进行,也可按主题类别进行。如能把每一篇独立的连续文本存为一个文档,在使用像 Wordsmith Tools 这种软件时将十分方便。词性赋码和语义赋码已在上

文中讨论过，这里不再赘述。

4）备检文件制作。每一个语料库都应该有自己的备检文件。备检文件中应包括对语料选择的依据的基本陈述、对赋码方案的制订原则的陈述和解释、语料库修订的历史记录以及扩充语料与原来语料的一致性保证。

5）语料库的发布和应用。可考虑通过CD-ROM或INTERNET发布建成的语料库。把语料库放置在服务器上，使用户通过客户端软件进行远程登录，实时索引；或通过WEB为用户提供实时索引服务。

综观现代学习者语料库建设，尚存在以下几个问题：1)语料来源和输入。我国的英语学习者群体庞大，但能够获得的学习者语言运用材料仅包括书面语，建库语料输入只能靠键盘输入，使建库过程十分艰苦，需投入大量人力物力。但资金的投入非常有限，制约了库容量的进一步扩展，也使开发建库过程更为复杂的学习者口语语料库建设力不从心。2)广大外语教师尚未完全理解语料库的意义，对语料库在外语教学中的应用缺乏必要的心理准备和技术培训，使得语料库这一强有力的工具难以推广。3)随着语料库技术的发展，需要开发针对中国学习者英语语料库的索引软件，这就需要外语研究者、教师以及计算机专家的通力合作。如今，我们仍然缺乏这种合作的基础。一方面，教师们需要功能更强、界面更友好、更容易使用且又能免费提供的共享软件。另一方面，开发这些软件以及为软件升级不断提供技术支持，都需要大量的资金投入。解决上述问题需要外语教育主管部门的支持，对具有长期效益的科研项目给予必要的支持。同时，广大外语教师应积极参与我国学习者语料库的开发和应用，消除所谓的“技术恐惧症”，积极交流科研成果，使语料库这一先进的科研工具为我国的外语教学作出更大的贡献。[16]

注 释

1 使用BNC语料库在线查询必须先登录BNC主页(*http://info.ox.ac.uk/bnc/index.html*),或通过匿名ftp(*ftp://sable.ox.ac.uk/pub/ota/BNC/SARA/*)下载语料库客户端工具软件SARA,注册后可以免费使用20天。COBUILD语料库可以通过登录伯明翰大学主页或COBUILD主页进行在线查询(*http://titania.cobuild.collins.co.uk/index.html*),但如果未经注册,只能使用其演示功能,每次索引只显示50行。TeCCon是英国翻译英语语料库索引软件,可通过登录英国Manchester大学理工学院主页(*http://ubatuba.ccl.umist.ac.uk/tec/index.html*)进行查询。该网址包含翻译英语语料库、索引软件以及对语料库和软件的说明文件。该网址提供两种方式进行在线索引,一是通过该网页上的索引程序插件进行索引,二是通过下载其客户端程序TEC Browser(*http://ubatuba.ccl.umist.ac.uk/tec/download/*)并将它安装在自己的计算机上后,再上网查询。

2 实际上,辛克莱本人并不否认直觉经验的价值。他认为,仅就直觉判断和假设与客观统计结果存在如此大的差异这一事实本身,就值得研究(1991:3-5)。

3 到2000年,JDEST的库容量已扩充为4,082,369个词。

4 有关LOB语料库的详细说明,可浏览 *http://www.hd.uib.no/icame/lob/lob-dir.htm#lob1* 相关文件。

5 本节有关语料库特征的基本观点得益于与导师杨惠中教授、卫乃兴博士、濮建忠博士以及雷秀云博士的讨论。

6 感谢薛学彦对该词进行的语料库分析。本例得益于与他的讨论。

7 共时语料库如 *The Complete Corpus of Old English*、*The Helsinki Corpus of English Texts*(*http://www.ling.upenn.edu/mideng/*)、*The Corpus of Early American English*、以及 *The Century of Prose Corpus*。参见(*http://www.lb.u-tokai.ac.jp/tono/index-f.html*)。

8 其他专用语料库如杨惠中教授等的"上海交通大学科技英语语料库"JDEST、*The American Heritage Intermediate* (AHI) *Corpus*(*http://ref.umdl.umich.edu/a/ahd/simple.html*)、*The Oxford Psycholinguistic Database* (*http://wapsy.psy.uwa.edu.au/uwa-mrc.htm*)等。

9 90 年代初，语料库在英国、美国、日本以及欧共体国家几乎是突然普及开来。里奇认为，对语料库方法的这种突然高涨的兴趣，“主要是由于语料库方法在具有商业价值的领域的广泛应用，诸如词典编纂、语音识别、语音合成以及机器翻译”(Leech 1991，注释 6)。语料库迅猛发展的另外一个主要因素是世界上对外语学习的极大需求，以及语料库在外语教学和学习(如“数据驱动学习”)的广泛应用。

10 Yukio Tono 在其语料库语言学网页上详细列出了目前世界上主要的语料库及链结网址(*http://www.lb.u-tokai.ac.jp/tono/index-f.html*)。

11 Concordance 1.1.3 由瓦特(R. Watt)开发。该软件可以对采用 COCOA标注系统的电子文本进行索引，提供搭配词信息。该软件最具特色的功能是能对索引自动生成网页在网上发布。在线注册需要付费。其试用版具备全部功能，可供免费下载，注册前可使用 30 天。下载网址：*http://www.rjcw.freeserve.co.uk/*。该软件一次处理语料量大约 16 万词左右。

12 这些学习者语料库的网址分别为：*Corpus of English by Japanese Learners*：*http://www.lb.u-tokai.ac.jp/icorpus/*；*Japanese Junior Highschool Students' Japanese/English Translation Corpus*：*http://www.ipc.hiroshima-u.ac.jp/~d052121/eigo1.html*；*Hungarian EFL Learner Corpus*：*http://ipisun.jpte.hu/~joe/jpu/info.html*；*Cambridge Learners' Corpus*：*http://www2.cup.cam.ac.uk/elt/reference/clc.htm*。

13 按照 COCOA 编码标准，标注索引信息放在尖括号内，括号内分两个部分，中间由空格分开，第一部分为索引信息，第二部分为分类信息。括号内信息可以是文本和数字。在标注时应注意使同一个标注放在一行内。在利用工具软件(如 *TACT*、*CONCORDANCE*)进行处理时，应根据所提供的出错信息反复修正文本中的错误格式，以保证处理结果准确无误。

14 该工具是利用 *MS WORD* 内含的“自动图文集”功能，并结合其菜单和工具条定制功能开发出来的。该工具条完成后，存为通用模板文件(*normal.dot*)。不同赋码者互相移植时，只需把该文件拷入..*office/template/* 目录下，替换原来的同名文件即可。使用时，从“视图→工具栏”中激活该工具。该工具可从本人主页下载：*http://corpusling.126.com*

15 在大学英语学习者语料库建设过程中，杨惠中教授要求我们对每次讨论所

提出的计划、方案、修订，以及建库过程中所遇到的问题和解决方案都详细地记录下来，作为备检文件。结果表明，这种做法对以后的工作帮助极大。本节对语料库建设的讨论有些是受杨惠中教授平时提出的一些思想和看法的启发，有些来自于我和我的合作者濮建忠博士在建库和数据处理过程中针对一些具体问题所进行的讨论和总结。

第三章　语料库证据支持的词语搭配研究

引　言

词语搭配(collocation)无疑是语料库语言学研究活动中最为活跃的研究领域,处于它的中心地位。这种局面主要由两方面的原因使然。从实践上讲,在语料库问世之前,人们不可能有效地获得大量的真实语言数据(authentic language data)对语言进行研究,而只好使用相当多的直觉数据(intuitive data)对语言进行描述。这种描述主要集中在抽象的句法结构和语义框架方面。基于现代计算机技术的语料库使得成百上千万词的连续文本可被有效地储存和检索,过去可望而不可及的词语行为(lexical behaviour)研究变得既可能又便捷。所以,词语行为 —— 主要是词语搭配的研究便成了极富活力的研究领域。从理论上讲,一系列重要的语言学观点无疑对词语搭配研究起了关键的指导和推动作用。这些观点主要包括下列内容:

第一,语言交际主要是词语驱动的(lexis-driven)。语言体系中的词语具有其意义潜势(meaning-potential)。人们的交际活动,主要是选择合适的词语,包括词语的适当语法形式,来实现意义的一个过程。从这个意义上说,传统语言学理论所建立的"句法第一,词语从属"的理论框架是靠不住的。这些语言学家试图使人们相信,句法是一套关于语言使用的普遍性的或整体的结构规则,词语仅仅是用以填充这个结构的材料。语料库语言学的研究表明,情况绝非如此。

第二,语言学研究应以词语研究为出发点。词语和语法的关系在语言使用中是一种共选关系(co-selection)。一定的词语和意义总是以一定的语法形式表现出来;一定的语法结构也总要以最经常和最典型的词语来实现。从某种意义上说,每个词都有其典型的语法结构,词的每个义项也有其一定的语法形式。在语言交际中,一旦词语被选定,相应的语法形式也就随之选定。也就是说,词语和语法是密不可分的。二者在语言使用中处于一个连续体(continuum)中,而词语搭配研究将词语和语法密切结合起来,正好研究这个连续体。在某种程度上,以词语为中心的语言学研究,是以将形式、意义和功能综合一体进行研究为特色的。

第三,有必要将语言的使用与运作看成是一个词语的网络系统。这个网络系统主要由词项(lexical item)、词组(phrases)、半固定词组(semi-fixed phrases)和词块(lexical chunks)构成。词项并不是按照句法规则的限制简单地组合在一起的,它总是服从于意义的实现与使用习惯,有规律地和其他词项结合成为词组、半固定词组和词块。而由此构成的结构可视为词语网络系统,或词语结构。当然,这是同一个硬币的另一面。但是,将语言运作视为一个词语结构有其独特的优势。它使研究者能更加系统、深入地研究词语行为规律,并进而研究语义、语用等方面的规律。词语结构与语法结构、语义结构和语用结构是相互依存,相互渗透的关系。在这一点上,词语搭配研究无疑是进行这些研究的一个合适的切入点,是最有意义和最为有效的一种语言学研究途径。

第四,语言学应研究真实使用中的语言。语言学家个人的直觉对语言学研究而言是极其有限的。为说明和论证某种理论框架而杜撰和生造的句子没有太大的效度,在很大程度上很可能会将人为的结构强加于实际语言运作,会对真实语言使用曲解。词语搭配之父弗斯(J. Firth)的重要语言学原则之一就是使用"验证数据"(attested data)(Firth 1957)。当代语料库语言学家辛克莱(J. Sinclair)则用"自然发生数据"(naturally occurring data)(Sinclair 1991)这一说法,其含义与精神实质与弗斯的主张是一脉相承的。

基于大量语料库证据的词语搭配研究正是这一学术思想的运用和体现。

由于上述理论和实践两方面的诸种因素，词语搭配研究成了语料库语言学领域里核心的研究内容。然而，关于词语搭配的界定，它的基本方法，词语搭配研究与一般语言学研究的关系等等，学术界正在进行各种各样的探索，尚没有定型的理论框架。本章着重对这些重要问题进行探讨，并对该领域里主要的观点与方法进行概略的介绍。

1 词语搭配的界定

1.1 广义的界定与狭义的界定

词语搭配研究之父弗斯曾说过，“由词之结伴可知其词”（“You shall know a word by the company it keeps”）（Firth 1957:12）。按照弗斯的观点，词语的结伴关系或共现关系（co-occurrence）是词语搭配的重要表现形式；词语搭配是一种意义方式（a mode of meaning）；习惯性搭配中的词项相互期待和预见（expect and predict each other）；词语搭配这种词语使用现象与类联接（colligation）有着内在联系（见卫乃兴博士学位论文：Towards Defining Collocations）。然而弗斯本人并没有对词语搭配进行严谨、系统的界定。语言学界对词语搭配的看法也纷纭复杂，研究方法也极不相同。在当今的语料库语言学研究中，词语搭配的界定可分为广义界定和狭义界定两种。前者以新弗斯学派的韩礼德（M. Halliday）和辛克莱为代表，后者以杰尔默（G. Kjellmer）等人为代表。

广义的词语搭配界定体系认为，词与词的共现关系是最重要的或者说惟一的标准。只要一个词与另一个词共现达到统计学上的显著程度，它们就构成搭配。这里，定量分析是最为重要的。词与

词之间的自由组合(free combination)和成语 (idiom)可能都被视为词语搭配。这些特点可由韩礼德提出的词语搭配定义看出。韩礼德认为,"词语似乎只需要线形共现,具有某种程度上的显著邻近,或者是在连续体上,或者至少是在某个截断点内。正是这种组合关系,被称之为搭配。"("Lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, either a scale or at least a cut-off point. It is this syntagmatic relation which is referred to as 'collocation'...")(Halliday 1976:25)。广义的词语搭配定义基本不考虑共现的词项之间的语法关系,也就是说,构成搭配的词不一定具有语法上的相互限制关系。例如,韩礼德(ibid:74)认为在下列两句话中:"argument"和 "strong"两个词就构成搭配,尽管二者分属于两个不同的句子,没有什么语法限制关系:

I wasn't altogether convinced by his argument. He had some strong points but they could all be met.

广义的词语搭配界定很难为多数语言学工作者接受。但这种界定是一种开放的研究理论体系,不受传统语言学理论中许多先入为主的观念束缚,借助大量的自然发生证据,可望发现语言运作的许多内容和机制,对语言使用作出新的描述与解释。放在新弗斯学派关于语言理论的整体框架中看,词语搭配是研究词语学(lexis)或词语结构(lexical structure)的一种重要手段和途径。这样就不难理解广义的定义框架了。

狭义的词语搭配定义主张搭配研究必须参照语法限制关系。这里有两方面的因素对搭配制约。其一是语法的限制作用(grammatical restrictions),其二是词语的决定作用(lexical determination)。其含义是,只有处于同一语法结构中的词项才可能构成搭配。然而处于同一语法结构中的词项不一定都是典型的搭配 ——这还要看词项之间共现的显著性程度。杰尔默将词语搭配界定为

"以等同形式超过一次重现、并且构成良好语法关系的词汇序列"("A collocation is a sequence of words that occur more than once in identical form (in the BROWN Corpus) and which is grammatically well-structured")(1987:133)。

该定义对词语搭配提出了一系列定量和定性标准。其中有些标准有待于商榷,但其最重要的标准之一无疑是"grammatically well-structured"一条。如果照杰尔默的标准来分析,在韩礼德列举的上述两句话中,"strong"和"argument"并不构成搭配,因为二者不存在语法上的限制关系,倒是"strong"和"points"可能是一种搭配,因为二者处于同一语法结构中。

狭义的词语搭配定义兼顾了词语行为中的词语因素(lexical factors)和语法因素(grammatical factors),既有定性的标准,又有定量的标准。它适用于基于小规模语料库证据进行的实用性研究,如旨在服务于 EFL 教学的搭配研究项目。然而,杰尔默提出的定义并非就是一个完美无缺的定义。事实上,词语搭配的定性分析与定量分析方面有许多问题需要讨论。词语搭配有其诸多的界定特点需要考虑。

1.2 词语搭配的界定特点

1.2.1 词语搭配:词在组合轴上的共现

语言学所研究的词语关系可分为聚合轴上的研究和组合轴上的研究。前者又称为纵向轴或垂直轴,后者又称为横向轴或水平轴。前者研究的是词语间可能存在的抽象的联想关系。这种抽象的联想向人们揭示词与词之间可能存在的同义、反义和上下义等关系以及由此可能导致的在句子中的相互替代。比如:

This paper describes the preliminary experiment...

在这句话中,*paper* 一词可能会被 *article* 取代,因为二者间存在着

·种聚合联想关系,是同义或近义词. 除 *article* 一词之外,*essay*、*thesis*、*research report*、*dissertation*、*research proposal* 等词语也和 *paper* 具有同样的同义或近义关系,因此也可能在这个句子中替代 *paper*。词语间的聚合关系一直是语义学研究的重要内容之一。聚合轴可以说是语言的意义潜势之所在。然而在语言交际中,意义的实现却是在组合轴上进行的。语篇以一种线型方式展开。词语间的这种线型关系即是词语搭配要研究的内容。如果说聚合轴上研究的是词语间可能存在的基于联想的某种关系,那么组合轴上研究的则是词语间实际上业已发生的组合关系。词语搭配研究的不是可能性的东西(what is possible),而是真实性的东西(what is real)和典型性的东西(what is typical)。

在语料库问世之前,要深入、详尽和有效地研究词语的组合关系是不太现实的。所以语义学家们大都着重探讨聚合关系,认为聚合关系的内容要比组合关系丰富得多,有意义得多(Lyons 1991:269)。今天,强大的语料库证据已经可供使用,研究词语间的组合关系已成为一种现实。而且,人们发现,组合关系具有丰富的形式和内容,组合轴上所实现的意义也十分丰富。仍以 *paper* 一词为例。在上海交通大学的 JDEST 语料库中,我们发现 *paper* 经常和 *discusses*、*describes*、*reports*、*presents*、*demonstrates*、*argues*、*introduces*、*addresses*、*examines*、*presents* 等动词共现,构成 N+V 搭配,表达一系列不同的意义。比如,*this paper describes* 往往用于学术论文的"引语"中或"摘要"中,向读者揭示论文所探讨的主要内容特点。下面是这个搭配的 14 行词语索引:

1	a long transform of length N.	Paper describes two techniques based
2	foam vaporization casting; the	paper describes application of the pr
3	hieving long term success. This	paper describes a TQM integrated educa
4	ontemporary methodologies. This	paper describes one of the most effec
5	ATION SYSTEM ABSTRACT This	paper describes the design and impleme
6	er,Maryland 20670 ABSTRACT This	paper describes the production plannin

7	n the line organization. This	paper describes a PEB method that has
8	oughout the winter season. This	paper describes the analysis performe
9	mes, Iowa 50011 ABSTRACT This	paper describes a set of computer assi
10	and municipa water users. This	paper describes the design, construct
11	y source of refluxing. This	paper describes an analysis of the wel
12	d Natural Gas AGR Process. This	paper describes the CNG process and ex
13	various functional areas. This	paper describes how the tools within
14	are briefly explained. The	paper describes the experimental techn

除了上述 N + V 搭配外，*paper* 还有其习惯性的 DET（+ ADJ）+ N 搭配。如 *this paper*、*the present paper*、*the previous paper*、*the preceding paper* 等等。其中 *this paper* 这个搭配颇能说明一些问题。传统的语言描述告诉我们，*this paper*、*that paper*、*the paper* 都是合乎语法的可能的用法。然而，JDEST 语料库证据显示，在学术英语中，*this paper* 是个使用频率极高的搭配，意为"本文"。经统计，在 *paper* 的有关使用实例中，*this paper* 占 76%，*the paper* 占 16%，而 *that paper* 则没有一例。实际上，*this paper* 基本上相当于 *the present paper* 的意思，是 *my paper* 或者 *our paper* 在学术英语中的典型说法。在整个 400 万词的 JDEST 语料库中，*my paper* 和 *our paper* 总共才出现 5 次，*that paper* 出现 2 次，其概率微乎其微。

paper 还和介词 *on* 等构成显著搭配，表达学术英语中各种各样的意义。这些情况说明，组合轴上的词语关系和它们所表达的意义远比语言学家所能想象的丰富。这些关系及其表达的意义正是词语搭配要研究的中心内容。只有当词项在组合轴上共现时，它们才有可能成为词语搭配研究的内容，这是语料库证据支撑的词语搭配的特点和生命所在。

然而，这并不是说所有在组合轴上共现的词，都构成实际意义上的搭配，虽然新弗斯学派这样认为有其道理。我们认为，在进行小规模的实用性研究时，仍有必要对组合轴上的词语进行定性的分

析,对它们进行区别。

1.2.2　词语搭配:组合轴上词语的有限组合

组合轴上的词语组合大致可分为三类,自由组合(free combinations)、有限组合(restricted combinations)和成语(idioms)。区分的标准有三个,搭配力(collocability)、结构可变性(structural mutability)和语义明晰性(semantic transparency)。搭配力指的是一个词与其他词共现的趋势或能力,当一个词与多个词共现并构成显著搭配时,它的搭配力就大,反之则小。结构可变性指的是词语组合的成分可以发生位移、被扩展和被别的词代替。意义明晰性是说词语组合的整体意义基本上可以清楚地由各个组成成份的字面意义看出。自由组合一般是一种结构松散的词语序列(lexical sequence),该序列的成分能较自由地和其他广泛的词项结合,构成类似序列。自由组合有着明显的意义明晰性,由各个成分的字面意义便可得知整个序列的意义。我们以动词 *buy* 所构成的 V + N 序列为例说明上述各点。语料库证据表明,动词 *buy* 可和表示各种各样意义的诸多名词构成 V + N 组合,如:

(1)

buy + a/the { house / flat / real estate / property }　(表示"房产"类的名词)

(2)

buy + a/the { dog / cat / eagle / parrot }　("宠物")

(3) buy {food / wine / alcohol / bread} ("食物")

(4) buy + a/the {machine / radio / TV / camera} ("设备与电器")

(5) buy + a/the {firm / factory / company / corporation} ("企业/组织")

甚至某些抽象意义的词也可作 *buy* 的宾语,与其构成 V + N 组合,如:

(6) buy + {freedom / balance / ideology / support} ("抽象概念")

上述例证均选自 BROWN 和 LOB 等语料库。它们说明,动词"buy"的搭配力很强,搭配范围(collocational range)很大,几乎是开放性的(openness);几乎所有的名词都可在 buy + N 组合中充当 N 的成分。这些组合大都有结构松散性或可变性的特点。如其中的 N 可以位移;*buy a book* 可变为 *select a book to buy*; *buy a firm* 中的 *firm* 可由 *company* 代替等。在意义上,整个序列的意义相当明晰,"buy"在所有这些组合中的意义基本不变。弗斯(1957)曾经说过,词语搭配中的各个成分相互限定、相互期待和相互预见。在由"buy"组成的上述组合中,各成分间的相互限定、相互期待和相互预见的关系比较弱和不太明显。由于这些特点,我

们把这类词语组合称为自由组合。

另一类词语组合是成语(idioms)。成语是一种在句法和语义上相当于一个单个词的词语组合,其结构相对固定,整体意义与各组成成分的字面意义无关。以 *kick the bucket* 为例。作为一个成语,其意义为"死亡",在句子功能与意义上相当于单个的词 *die*。它所表示的意义"死亡"显然不是 *kick*、*the* 和 *bucket* 各成分的字面意义相互作用的结果。也就是说,它的意义是不透明的(opaque)。它在结构上是相对固定的。考伊(A. Cowie)和麦金(R. Mackin 1975)认为,*kick the bucket* 中的 *kick* 一词可以用于几种不同的时态,如:

The old man kicked the bucket last night.
The old man has kicked the bucket.
The children have no love for the old man; they will be pleased if he kicks the bucket tomorrow.

但是该成语几乎从不用于进行体中

* The old man is kicking the bucket. (*表示该句不可接受)

如果用于强调句 * *What the old man did is kick the bucket*.中,它所表示的意义就完全不同了。它同样也不能变为被动语态 * *The bucket was kicked*.

成语的成分是不能随便用其他词替代的。*kick the bucket* 中的 *bucket* 不能换为 *pail* 等。就搭配范围而言,成语组成成分的搭配力已减小到了极点。

介于自由组合与成语间的是一种有限组合。与自由组合和成语相比,有限组合有其明显特点。在自由组合里,词的搭配范围几乎是

开放性的，而构成有限组合的词项的搭配范围却相对较为狭小；成语基本上是一种固定不变的词语组合，而构成有限组合的词项则有其一定的自由度和其他词项结合。我们认为，"搭配"这个术语意味着词的搭配力是受方方面面因素限制的，因而搭配范围是相对狭小的，介于开放的自由组合与几乎封闭的成语间。因此，应当把，也只能把有限组合看做词语搭配。以名词 *proposal* 为例。英国伯明翰大学的 COBUILD 语料库证据显示，与 *proposal* 搭配的动词有 *make*、*rejected*、*put forward*、*support*、*accepted* 等十多个词语左右。下面是各搭配词的统计数据，表中列出了各词在语料库中出现的总频数，与 "proposal"共现的频数，以及计算出来的 T 分值。

表 3.1　有关 *proposal* 的搭配词及其统计数据

Collocates	Total occurrences	Co-occurrences with node	T-score
rejected	2227	49	6.912 898
put forward	30 919	53	6.117 334
support	14 095	33	5.072 799
accepted	3543	25	4.805 997
backed	2377	12	3.271 441
submitted	572	11	3.269 407
grant	3050	11	3.064 805
approved	1894	10	2.998 299
opposed	1804	7	2.835 364
veto	429	7	2.601 358

当然，表中数据并没有将能和 *proposal* 一词搭配的词穷尽。由于语料库的某些特点所致，*raise a proposal*、*object to a proposal* 等没有反映出来。即便如此，我们也可以看出 *proposal* 一词的搭配范围是有限的。

英语中的许多词语组合都是有限组合，也就是通常所说的词语搭配，如修饰语—中心词类：

a broad hint, a heavy smoker, an overwhelming majority, constructive criticism, concerted efforts, flawless English, unfailing support, resounding success, sweeping statements

动词—补足语类：

address an issue, adopt a policy, build a reputation, draw a conclusion, earn an academic degree, fight corruption, present a paper, resolve the differences, mete out a penalty

就语义特点而言，词语搭配的意义是明晰的。但搭配伙伴间语义上的相互限定关系较强。这种相互限定关系对搭配的整体意义产生一种影响；它不仅仅是各搭配伙伴字面意义的简单相加。这种较强的意义相互限定关系使搭配伙伴能够相互期待和相互预见。在结构上，词语搭配容忍一定程度的替代、位移和扩展。比如 *supported the proposal* 中的 *supported* 可替换为 *backed*，说成 *backed the proposal*；*draw a conclusion* 中的 *draw* 可替换为 *reach*，*come to*，说成 *reach a conclusion*，*come to a conclusion* 等。但是，替换是受搭配范围制约的，超出搭配范围的替代会导致不合习惯或不正确的说法。

词的搭配范围具有有限性，这一点是和语言使用的典型性(typicality)密切相联的。英国语料库语言学家汉克斯(P. Hanks)曾对语言使用中的词语选择与典型性的关系评论道，"选择的根基存在于典型性这个概念之中，英语的词汇并不是自由地和任意地结合和再结合的。不但典型的语法结构和形式类别可以被观察，而且典型的搭配词也可观察。"("The basis of choice has its root in the notion of typicality. The words of English simply do not, typically, combine and recombine freely and randomly. Not

only can typical grammatical structures and form classes be observed, but also typical collocates")(1988:121)。

词语搭配具有典型性正是由于词的搭配力和搭配范围有限。这种有限性和语言使用的地道性(idiomaticity)与自然性(naturalness)密切相关。在使用词语搭配时违反有限性,就意味着破坏地道性和自然性。在外语学习中,词汇学习的主要内容就是学习词的有限组合。不掌握词的有限组合而只记忆词的"意义"是没有多大价值的。在很大程度上,能否较熟练地使用外语要取决于能否熟练地使用典型的词语搭配。

不过,自由组合、有限组合与成语三类词语组合间并没有截然分明的界限。三者在三条区分标准——搭配范围、结构可变性与意义明晰性上只是程度上相对不同而已。就大多数词组而言,人们很难说它们在结构上是绝对固定不变的或绝对可变的,在意义上是绝对明晰的或不透明的。在搭配范围上也是如此,人们很难说一个词的搭配范围包括了几个词项才能算是开放的或狭小有限的。所以,这些区分的尺度只是一些定性的分析,不宜作量化处理。三类词语组合最好被看做是处于一个连续体(continuum)中,自由组合和成语各位于连续体的两端,而有限组合则位于两者中间的某个地方。在连续体上,三者间不可避免地有一些相互重合的部分,如下图所示:



图 3.1 词语组合的连续体

在下面的几节里,我们将分别讨论词语搭配的其它几个重要界定特点。它们分别是:词语搭配的因循性、意义—形式限制性、反复重现性与统计度量性、语域相关性和长度多样性等。

1.2.3 词语搭配:语言使用的因循性

词语搭配在很大程度上是一种习惯性或常规性的词语行为方式,这就是它的因循性(conventionality)。因循性可以从三个方面来探讨,即社会文化的因循习惯、学术文化的因循习惯和语言内部运作的因循习惯。首先,词语搭配方式直接受社会文化方式的影响。一个社会的文化传统、生活方式、风俗习惯和人们的思维模式不可避免地会反映到语言中来,词语搭配在一定程度上受其制约。比如,英国人有下午喝茶的传统习惯,因此英语里就有 *afternoon tea* 这个搭配。在 COBUILD 语料库里,*afternoon tea* 出现的频率相当高,这里显示 16 行词语索引供说明:

1.	in Club World when featured,	afternoon tea and light snacks are available at
2.	83424). Double room, dinner,	afternoon tea and breakfast from * 92 per
3.	at a pace, to leisurely	afternoon tea and elegant dinner parties. The
4.	to invite your friends for	afternoon tea or after-dinner coffee, or Sunday
5.	reakfast and lunch on deck.	Afternoon tea and late-night buffet. [c] diagram
6.	past four. I know I had	afternoon tea as soon as I got in." [p] I see,"I
7.	in true English style with	afternoon tea at The Ritz. Here you will be
8.	The event will begin with	afternoon tea at 3pm followed by the show which
9.	meal, morning coffee or	afternoon tea; don't worry if you are not staying
10.	of 12, would go nicely with	afternoon tea. Farmhouse country cake at Sainsbury
11.	stretch to lunch and even	afternoon tea; Hill was a great and generous
12.	unparalleled music industry	Afternoon tea in grand hotels Scottish tweed and
13.	ire in the Great Hall while	afternoon tea is being served. [p] Many guests
14.	each room. [p] A delightful	afternoon tea is served daily in the hotel's
15.	some of his men for Sunday	afternoon tea. It was my [10] duty to welcome them
16.	House of Commons, including	afternoon tea with Tony Blair MP, Leader of the

相比之下,早上喝茶却不是多数社会成员的习惯。所以,英语中 *morning tea* 的说法就少得多。同样由于这个喝茶的习惯,我们会经常在英语中见到 *tea break*、*tea service*、*prepare the tea* 和 *tea or*

coffee 等习惯性搭配。另外，由于心理习惯和看问题的角度不同，英国人用 *black tea* 来表示汉语的“红茶”这个意思，用 *strong tea* 表示“浓茶”这个意思。文化生活对词语搭配的影响和限制在许多方面都有体现。在英国，律师这个职业是和 *bar* 联系在一起的，牧师是要穿 *frock* 的，等等。因此，律师和牧师被解雇要分别用到 *disbarred* 和 *unfrocked*，其他不同职业的人被“解雇”、“辞退”或“开除”则用其他不同的词语组合来表示：

Barristers are disbarred.
Doctors are struck off.
Solicitors are struck off the rolls.
Officers are cashiered.
Priests are unfrocked.
Schoolboys are expelled.
Students are sent down.
Footballers are suspended.
Working men are sacked
Chairman of regional gas boards is sent on indefinite leave.

(Mitchell 1975:122)

第二，词语搭配是学术文化习惯的反映。学术文化指的是科学研究领域里长期以来建立和发展起来的并为科技和学术研究人员所遵循的专业用语习惯：专业英语的用语特点与普通英语有很大不同；甚至各学科领域里的用语特点也不尽相同。我们在前面的1.2.1中看到，在学术英语中要表达“本文”、“该研究”等意义要 *this paper*、*the present paper*、*the present study* 等搭配，而不用 *the paper*、*my paper*、*our study* 等。虽然后几种用法也是符合英语语法的表达方式，但却不符合学术论文的写作惯例。可以毫不夸张地说，科学研究的每个领域都有自己的学术文化习惯以及由此而产

生的词语搭配,我们将在 1.2.6 中详细讨论这个问题。

第三,词语搭配是语言内部运作机制的一种因循习惯。语言符号的能指和所指之间,词语搭配的方式及其表达的意义之间具有很大程度上的任意性。在普通英语中,联合国安理会的"常任理事国"要用 *permanent members* 这个搭配表达。为什么不用 * *perpetual members*, * *permanent constituents* 等来表达呢?因为它们不是惯例性的词语搭配方式。一个词语搭配经使用后,为大多数语言使用者接受、模仿和使用,就形成了惯例,被不断反复使用。英国语言学家米切尔(T. Mitchell)曾生动形象地用 *work* 一词来说明词语搭配的因循性特点。他说,"不同职业的人员做不同的工作要用不同的词语搭配来表达:水泥厂工人在水泥厂 *work*(工作),其他不同职业的人可能对艺术品 *work*(进行创作);但是,神职人员"做善事"用 *perform good works* 表达,"建造水泥厂"用 *build cement works*,"创作艺术品"用 *produce works of art* 等,这些都是词语搭配的因循性"("... Men — specifically cement workers — work in cement works; others of different occupation work on works of art; others again, or both, perform good works. Not only are good works performed but cement works are built and works of art are produced"(1975: 117)。

词语搭配的因循性,尤其表现在一些已经丧失词语内容的词(delexicalized words)的搭配行为上。英语中有许多这样的虚化词项,它们本身几乎不再有什么明确的词语意义;其意义已嵌入其搭配伙伴中,如 *make*、*do*、*take*、*have*、*get* 等词。这些词的搭配行为似乎在受语言运作机制的调节和制约;它们各有其较为固定或半固定的搭配伙伴,与其组成一种已经高度惯例化的词组;我们不能随便将词组中的虚化词项换为别的虚化词项。以动词 *make* 为例,*make* 经常与下列名词搭配使用:

make + an/a + { experiment / telephone call / trip/journey / visit / enquiry / claim / decision / suggestion }

在上述搭配中，*make* 基本上已没有自己明确的意义，它的意义体现在与其搭配的名词中。我们可以较有把握地说，*make an experiment* = experiment; *make a telephone call* = call; *make an enquiry* = enquire; *make a claim* = claim; *make a decision* = decide; *make a suggestion* = suggest。在这里，*make* 与各搭配伙伴组成了一种已经高度惯例化的词组；我们不能随便将其换为别的虚化词项。比如，我们可以说 *make an experiment* 或 *do an experiment*，但却不说 * *take an experiment*；我们说 *make a telephone call*，但却不说 * *do a telephone call*；我们说 *make a visit*，但却不说 * *do a visit*、* *take a visit*，等等。在语言使用中，在多大程度上能将这些词组中的虚化词项替代为另一个虚化词项，替代为哪个词，都要遵循词语搭配的因循性。本族语者对词语搭配的使用是因循性的行为，第二语言学习者学习词语搭配时也必须依照惯例。

1.2.4 词语搭配：意义—形式的综合体

语言交际中的一切形式选择都是为了实现意义，词语搭配也是如此。从前面几节内容，我们可以看到，一个词和不同的词搭配，就表达不同的意义。比如在 1.2.2 中讨论的 *proposal* 一词的搭配词：虽然该词的搭配范围有十多个动词，但每一个动词和 *proposal* 组成的搭配都表达一个具体的意义。在这一节里，我们从语言交际意义的宏观体系上去探讨搭配所实现的意义。按照韩礼德的体系，

意义在宏观上可分为三类，概念意义(ideational meaning)、人际意义(interpersonal meaning)和篇章意义(textual meaning)。概念意义表述的是我们在客观世界里的经历以及我们内心世界里的想象。人际意义表述的是交际双方用语言做事，使交际得以进行下去的活动。篇章意义直接关系于语篇信息的安排和组织(1985：53)。词语搭配同样表达这三种意义。我们以 *experiment* 一词的有关搭配来说明这个问题。在学术英语中，与 *experiment* 搭配的动词主要有下列几个：

performed
conducted
carried out
presented
described
controlled
} an experiment

这些搭配实现的是不同的概念意义，表述了科技界常常进行的活动、实践和经历：进行实验，开展实验，报告实验，描述实验，控制实验等。然而，*experiment* 还常和其他一些动词搭配，构成另一类不同结构的搭配：

experiment/s {
demonstrate/s
show/s
indicate/s
suggest/s

这一类搭配实现的是篇章意义，常用来陈述作者的某种观点和评论，如：

New laboratory experiments suggest that noble gases

trapped within the icy nuclei of comets may account for the abundance of xenon in the atmospheres of Earth and Mars.(陈述观点)

或者用来报告某种结果和事实:

Experiments show that the application of the uniform direct stress does not produce any sheer distortion of the material.(报告结果和事实)

在这种情况下,词语搭配起着组织篇章信息、预示篇章发展的重要作用。

上面说的是词语搭配实现的意义。一定的语言形式实现一定的意义;一定的意义以一定的语言形式为载体。我们说词语搭配是意义—形式的综合体,就是说它的意义与形式不可分割,二者融合为一个统一的整体。在上述例子中,当 *experiment* 构成的搭配实现概念意义,表述科技界和学术界的经历和活动时,动词多用一般过去时形式,*performed*、*conducted*、*carried out* 等等。当搭配实现篇章意义,组织篇章信息,预示篇章发展时,动词多用一般现在时形式,如用 *demonstrate/s*、*show/s*、*indicate/s*、*suggest/s* 等等来陈述作者的观点或报告结果和事实。一般现在时的使用有强调"与目前相关"的功效,对组织篇章信息有较好的作用。词语搭配的形式和意义都融合在一起。

这就是说,研究词语搭配,必须考虑有关搭配词的语法形式。事实上,语法形式对词语搭配有着一定的限制作用。这种语法限制作用(grammatical restrictions)可用韩礼德的几个例子来说明。在英语中 *strong argument* 是个常见的词语搭配。在不同的语法结构中,该搭配会变成不同形式的搭配(Halliday 1976:74),如下列例句所示:

He argues strongly.

He put forward a strong argument for it.

I don't deny the strength of his argument.

His argument was strengthened by other factors.

上述例句中,*argue strongly*、*strong argument*、*the strength of his argument*、*argument was strengthened* 都是合乎语法的搭配(well-structured collocations)。在这些搭配中,搭配伙伴的形式直接受语法因素的制约。

从语言运作的更高层次上讲,语法必然对词语搭配具有限制作用。既然词语搭配是组合轴上词语的结合,那么它必然受诸多组合因素(syntagmatic factors)的制约。组合因素之一就是语法,或者说句法。弗斯在阐述词语搭配时提出了类联接(colligation)这个概念(见刘润清等 1988:99)。米切尔(1975:120-122)对类联接与搭配的关系阐述得较为清楚:类联接指的是语法范畴的结合。类联接不是与词语搭配平行的抽象,而是高一级的抽象。类联接是关于句法结构和范畴的陈述,搭配则是类联接的具体实现。

因此,严格意义上的词语搭配应当是指位于同一语法结构内的词与词之间的组合。那么,词语搭配的界定必须限定在类联接的框架内。类联接就是词语搭配出现于其中的语法结构和框架。一个类联接代表了一个类别的词语搭配,可称为搭配类(collocational class, collocational type)。比如,N + V 就是一个类联接,它代表一类搭配,在研究中常称之为 N + V 搭配。*Data shows*、*heart throbs* 和 *water evaporates* 等等都是 N + V 搭配的具体实例。V + N 也是一个最常用的类联接,或搭配类,*commit suicide*、*launch a project*、*pursue a study*、*perform a test* 等则是其具体实例。

类联接多种多样,有繁有简。a + n + of 也是一个类联接,也代表一类搭配,*a sort of*、*a pair of*、*a couple of*、*a series of* 等等是其具体的实例。too + adj. + to 也是一个类联接和代表一类搭配,*too good to*、*too hard to*、*too easy to* 等等是其具体的实例

(Renouf and Sinclair 1991)。这些情况属于较为复杂的类联接。类联接界定多少,繁简到什么程度,要视具体的研究项目及其目的而定。就一般的词语搭配研究而言,类联接的界定限在词类的层次上,大约有十余个,代表十余种词语搭配。下表显示的是一般性研究项目中常见的最简单的类联接形式,或词语搭配类。

表 3.2 常用的词语搭配类联接

Types	Instances
1. N + N	quality improvement, death toll, price cut, team work
2. ADJ + N	constructive criticism, statistical evidence, empirical study
3. N + PREP	argument against, awareness of, evidence for, paper on
4. N + PREP + N	issue in management, varieties of data, study of the process
5. V + N	perform an operation, earn a degree, settle the dispute
6. V + ADV	judge harshly, refuse flatly, fight desperately, draw heavily
7. ADV + V	highly recommend, seriously affect, persuasively suggest
8. V + PREP	apply for, adhere to, alternate with, disassociate from
9. ADV + ADJ	sparsely populated, utterly helpless, diametrically opposed
10. PREP + N	in question, in value, by weight, under discussion
11. N + V	heart ...sinks, data... shows, tide..recedes, water...evaporates
12. phrasal verbs	take off, break down, point out, do away with, make believe

将词语搭配研究限定在类联接,也就是语法框架之内进行,并不是要将现成的语法体系和模式强加于语料库证据,也不是片面地

强调句法对词语行为的限制作用。词语和语法是相互影响的。从另一方面讲，语法不是空中楼阁，不能凭空杜撰。必须有大量的词语搭配实例作依据，才能概括、确定一个类联接。在具体的研究中，传统的语法结构描述可供参考，但更重要的是依照语料库证据来概括和确定类联接或搭配类。比如，在学术英语中，V-ing 类的形容词和 V-ed 类的形容词修饰名词的实例很多，如 *increasing amount/s*、*increased amount/s* 等等。因此，在研究学术英语的词语搭配时，就应当考虑建立 V-ing + N 和 V-ed + N 这两个类联接了（参见本书第六章）。

1.2.5 词语搭配：词语的反复共现及其统计度量

前面，我们谈到词语搭配研究的是语言使用中词项的共现现象。然而，并不是所有一经共现的两个或几个词，都可以说是词语搭配。只有在文本中反复共现的词，才可是语料库研究中一般意义上的搭配，因为词语搭配揭示的是语言使用的典型性（typicality）和地道性（idiomaticity）。只有反复共现的词语间才可以说具有搭配关系，才反映了语言使用的典型性和地道性。事实上，自由组合与有限组合的区别之一就是前者是语言使用者的一种临时性行为，不可能在文本中反复出现，后者是语言社团的习惯性行为，在文本中反复出现。仍以与 *proposal* 一词搭配的动词为例。我们之所以将 *put forward a proposal*、*rejected a proposal*、*accepted a proposal* 视为搭配，是因为它们在 COBUILD 语料库中反复出现。然而 *torpedo a proposal* 在该语料库中虽也出现了，但频率很低，不足以证明其为搭配。所以，*torpedo a proposal* 基本上是一种自由组合，一种为增强语言使用新颖效果的比喻性用法。

这就涉及词语搭配研究中的统计度量问题。语料库研究的重要手段之一就是其量化分析。量化分析基于各种统计度量方法（详细参见第四章、第五章）进行。在对词语共现进行量化分析时，常用的统计方法有 MI 值，Z-分值与 T-分值（详细参见本书第四章、第五章）。

反复共现这个界定标准及其统计手段是基于语料库证据的词语搭配研究的重要标准和特点。在语料库中仅仅出现一次的语言现象没有太大的研究价值和意义。所以，语料库研究者注意的焦点是那些在语料库中共现达一定次数的词语。反复共现这个界定标准是一种客观的标准、数据驱动的标准。它使得研究者从证据出发，在量化分析的基础上，对词语搭配的情况进行概括和描述。很多在真实语言使用中经常共现的词语都可以被计算机捕捉和发现。这一点是过去基于直觉的研究难以做到的。比如，*strong evidence* 和 *hard evidence* 都是英语中较为常用的词语搭配。但是 JDEST 语料库的证据显示，在学术英语中，*empirical evidence*、*experimental evidence*、*strong evidence* 是反复共现的词语搭配，但“hard evidence”却不是，如下面的索引所示：

1	ersals constitutes the only empirical	evidence for the idea that linguistic structure
2	I There is little reliable empirical	evidence upon which to base any thoroughgoin
3	discuss below, the positive empirical	evidence needs to be interpreted more care
4	vement. As we sought clear empirical	evidence that time-management practices influen
5	s one of the doctrines with empirical	evidence that highlights it. SENSORY AND PERCEP
6	ls. There is considerable experimental	evidence pointing to a relationship between inc
7	ans. Conversely, there is experimental	evidence that the profile of the action potenti
8	erability, and to present experimental	evidence that supports this model. SLUDGE FREE
9	e below we summarize more experimental	evidence which can verify to the point. Prima
10	t atrial muscle. There is experimental	evidence that the activation of the myocardial
11	y the feces of the adult fleas. Strong	evidence suggests that fleas descended from a w
12	ferentiating neurons pro- vides strong	evidence that translation occurs in these neuro
13	ly evident. Thus, there is very strong	evidence that diffusion accompanies successful
14	ough it is difficult to provide strong	evidence for some reasons related to learner
15	gure 6.5). However, there is no strong	evidence to suggest they form complexes. Quite

在这种情况下，研究者可以将上述几个反复出现的词语组合描述为学术英语中较为常用的词语搭配，从而勾画出该领域里语言使

用的一些特点。

1.2.6　词语搭配：不同语域里词语使用特点的反映

词语搭配明显地受语域(register)的影响。弗斯语言学理论中的一个重要概念就是语境。韩礼德将语境的概念发展成为语域；语域又被系统地分述为域(field)、旨(tenor)和式(mode)(1978:31－33)。简单来说，域指语言使用的习惯性环境，比如某个专业活动领域；旨指的是语言交际者之间的角色关系与交际的目的；式指的是交际的渠道或媒介，如口语和书面语。任何语言使用都受到语域因素的影响和制约，不同语域的语言使用呈现着不同的特点和规律，词语搭配也是如此。在应用语言学和语言教学领域，*student performance*、*discourse develops...*、*involve learners into interactive classroom activities*、*backwash effect of test* 等词语搭配是十分普遍的。英国的房地产商在做商业广告时常用 *delightful residence*、*tastefully modernized*、*having the benefit of double glazing* 等词语搭配手段。关于资本主义国家劳资关系的新闻报道中会常用 *stage industrial action*、*propose redundancies* 和 *threaten cuts* 等词语搭配(McCarthy 1995:64)。

在一种语域里使用频率极高的词语搭配，在另一种语域里可能很少使用，或者几乎不出现。比如，在医学领域里，如要表达"某种疾病的患者"和"有某种症状的患者"这些意义，医生和医学研究人员通常用 patient/s ＋ with ＋ N 这种搭配。其中的 N 是表示具体疾病的名词，如 *breast cancer*、*liver disease*、*leukemia* 等；或者是表示具体症状的名词，如 *chronic HCV*、*syndrome*、*bacterial sepsis* 等。在 JDEST 语料库中，*patients* 一词总共出现了 948 次，其中 238 次是用在上述形式中，如下列索引所示：

1	). Moreover, studies of tumors from	patients with both breast and ovarian cancer have
2	itivity (67 vs 73%). Most series of	patients with chronic HCV, including patients eval
3	if they resided in a nursing home.	Patients with cognitive :mpaiment, as determined

4	We have found that 85 per cent of	patients with a femoral artery areurysm have some
5	time frame of a few minutes. Most	patients with a pheochromocytoma have markedly el
6	le for the higher mortality seen in	patients with acalculous cholecystitis, although c
7	y described autoimmune phenomena in	patients with acquired immune deficiency syndrome(
8	s are almost invariably present in	patients with active disease. The most common is
9	lstone pancreatitis is suspected if	patients with acute pancreatitis present with bil
10	es in marrow transplantation, many	patients with acute leukemia can be cured. These t
11	e levels should be obtained in all	patients with acute pancreatitis because of potent
12	actable form of PCT occurs in some	patients with advanced renal disease. Associations
13	the severity of the pancreatitis.	Patients with alcoholic pancreatitis tend to have
14	f these problems are diverse. In	patients with alcoholic liver disease, bleeding ca
15	sponse to induction chemotherapy in	patients with AML, independent of disease category

然而在 BROWN 语料库和 LOB 语料库中，这种搭配仅仅各出现了一次：

Brown: B13 0500 Several dentists and patients with special dental problems have ...

LOB: G O Anticoagulant may arise in patients with haemophilia and Christ ...

词语搭配的这种语域制约性是其界定体系中的一个重要参数。研究者在观察与分析搭配现象时，要具有"语域"意识。许多词语共现现象在普通英语里似乎没有太大的价值与意义，但在某一个具体活动领域可以构成一种重要的搭配方式，表达重要的概念和实践活动，或实现重要的语篇意义。有些词语搭配在普通英语里极为普遍，但在学术英语中却不常用，或者用另一种表达方式。这些都是学术文化的因循制约作用。

另一个问题是词语搭配的长度。从上面讨论的搭配实例可知，词语搭配反映的可能是两个词之间的共现关系，如 *supporting evidence*，也可能是几个词之间的共现关系，如 *this paper describes*，

也可能长至整个词组或分句,如 *too good to be true* 等。这些不同的长度可概括为“一个词语序列”(a lexical sequence, a string of lexical items)。

综上所述,词语搭配的界定特点可归纳为这么几条:

1. 词语搭配是组合轴上融意义与形式于一体的词项共现
2. 词语搭配是因循性的语言使用
3. 词语搭配是文本中反复重现的词项结合方式
4. 词语搭配在很大程度上受语域的影响或制约
5. 词语搭配是一个词语序列的重现
6. 词项间相互搭配的关系可由统计度量确定

上面这些界定特点可综合起来,从而得出关于词语搭配的工作定义。用一句话表达就是:词语搭配是在文本中为实现一定的意义从而以一定的语法形式因循组合使用的一个词语序列,构成该序列的词语相互预期,以大于偶然的机率共现。(A collocation is a conventional syntagmatic association of a string of lexical items which co-occur in a grammatical construct with mutual expectancy greater than chance as realization of meaning in texts.)

定义中的 conventional 一词明确地表达了词语搭配的因循性性质; syntagmatic association 界定了词语搭配的组合属性;a string of lexical items 揭示了搭配单位是个序列,可长可短; mutual expectancy greater than chance 表明了词语搭配的反复重现性及其统计度量性; as realization of meaning 和 syntagmatic association 相结合说的是词语搭配的意义——形式限制性,而 in a grammatical construct 则明确地表述了词语搭配的语法限制性; texts 一词揭示了词语搭配的语域限制性特点,因为任何文本都属于一定的语域。

另一个尚待澄清的问题是搭配研究的中心内容。如果说词语搭配研究的是词项在组合轴上反复共现的情况。那么,这些反复共现的词项可能既有开放性的实词(open class words, content

words)，也有封闭性的功能词(closed class words，function words)。前者又叫词汇词(lexical words)，后者又叫语法词(grammatical words)。在传统的语言描述中，这两类词的切分是泾渭分明的：前者是表示"事物"、"性质"、"状态"或"行为"等具体意义的词，主要有名词、动词、形容词和副词，如 *book*、*run*、*musical*、*quickly*；后者指那些主要表达语法关系和在句子中起连接作用的词，主要有连词、介词、冠词、代词、助动词，如 *and*、*to*、*the*、*his*、*do* 等。这种划分在总体上是合乎实情的。词语搭配研究的重点应是实词之间的共现行为。因此，在定义中，我们用 lexical items 这一说法，以区别于 grammatical words。功能词虽然也有其典型的搭配行为，但这些搭配行为主要反映语法结构等，似不应成为研究的重点。在杰尔默的定义中(见 1.1)，功能词和实词是不作区分的，所以，在他的研究中，*must consider*、*will reduce*、*will prevent*、*an individual*、*our results* 等都是词语搭配(1991:116－120)的内容。我们认为，这些并不是词语搭配要研究的中心内容。但是，这并不是说功能词的搭配行为完全不应研究。从另一方面讲，功能词和实词间的界线并不是截然分明的。在语言交际实践中，人们对有些功能词的选择并不是基于它的语法连接作用，而是根据其所表达的词语意义。辛克莱(Jones & Sinclair 1974:30)曾举过这样的例子来说明该问题：*a house in London* 和 *a house near London* 两个词组的意义区别就是因为两个介词 *in* 和 *near* 的意义差别所致。所以他认为，语法与词语是一种"相互渗透"的关系，(1966:411)。韩礼德也认为："所有形式的词项都类属于两种范型。从语法的角度描述其所进入的封闭系统和有序结构时，它们是语法词。从词语的角度描述其所进入的开放系统和线性搭配时，它们是词语词"("All formal items enter into patterns of both kinds. They are grammatical items when described grammatically, as entering(via classes) into closed systems and ordered structures, and lexical items when described lexically, as entering into open sets and linear collocations")(1966:155)。

事实上，有些功能词组成的搭配表达了丰富的语篇意义，比如 *We will first discuss* ... 这样一个片语。该片语在学术论文中用得极为广泛。它的主要作用是起语篇组织作用，向读者揭示论文的篇章结构。我们不能因为 *will* 是所谓表示语法概念的功能词，而将该片语排斥在搭配研究的范围之外。在这种情况下，我们完全可以将该片语作为一个语篇连接语加以研究。这就是说，我们认为词语搭配研究的中心内容应是实词在文本中的行为。但在实词和功能词的区分上，我们不主张截然区分，一切应以词语在文本中实现的意义为依据来判断取舍。

2　基本路径与方法

2.1　提取节点词在语料库中的所有搭配词

基于语料库证据的词语搭配研究所采取的基本方法，是先从语料库中将与某个词共现的所有词语提取出来，然后计算各个共现关系是否显著，以确定其在多大程度上反映了词语间的相互预见关系，也就是说在多大程度上构成了搭配。为此，语料库语言学家建立了一整套的概念、思想和技术方法，包括节点词（node）、跨距（span）和搭配词（collocate）。节点词又叫搜索词（search word）或关键词（key word）。人们每一次在语料库中查询，都要将自己感兴趣的某个词从键盘上输入计算机，计算机则按照编好的程序，用 KWIC (key word in context) 的方法显示出词语索引。在每一行词语索引中，关键词，也就是节点词或搜索词，总是居中出现，而左右则是构成其语境的词语。语料库中的每个词都可以是节点词。选取哪些词为节点词完全由研究者根据其研究内容和研究目的而定。跨距指的是节点词左右的语境，以词为单位计算，不包括标点符号。假如将跨距定为 -4/+4，意思是说在节点词左右各取

4 个词为其语境。比如在 *For many years linguists have tried to define collocation in various ways*. 中,如果取 *define* 为节点词,那么,它左边的 4 个词,包括 *linguists*、*have*、*tried*、*to*,和右边的 4 个词,包括 *collocation*、*in*、*various*、*ways*,就共同构成了它的语境,也就是所说的跨距。与跨距相关的是距位(span position)。距位说的是跨距内每一个词所处的相对位置,常用 N-1,N+1,N-2,N+2 等表示。在上句话中,*linguists* 位于 N-4 距位上,*ways* 位于 N+4 距位上;*have* 位于 N-3 距位上,*various* 位于 N+3 距位上,等等。所有落入跨距内的词都是与节点词共现的词项。其中,有些共现词与节点词具有句法限制关系,也具有统计上的显著意义,就是研究者所要的搭配词。根据搭配词所处的相对位置,研究者又常用"左搭配词"(left-collocates)指位于节点词左边的搭配词,用"右搭配词"(right-collocates)指位于节点词右边的搭配词。

这种方法的基本思想是提取节点词的 2SN 个共现词,用于观察和研究。其中 2S 代表跨距,N 代表节点词在语料库出现的总频数。节点词在语料库中出现了 N 次,就意味着查询时要有 N 行词语索引出现,每一行中有 2S 个搭配词被提取,那么,节点词在语料库中的所有共现词就是 2SN 个。

2.2 跨距的长度

跨距长度的界定直接关系于搭配词提取的结果。词与词相距多远仍可能具有搭配关系?研究者对这个问题的看法也直接影响跨距的界定。一般来说,一个词与节点词的距离越远,它和节点词之间的相互吸引力,或者说相互预见力就越弱。但是情况并不总是这样。有些常见的词语搭配往往是习惯性地被分隔而处于非连续的位置上。请看下例中 *look up* 这个词组动词的使用情况:

在 She was obliged to look up the unknown word in the dictionary.

在此句中，*look up* 属于连续性搭配。而在很多情况下，这句话又可说成：

She was obliged to look the unknown word up in the dictionary.

在这种情况下，*look* 和 *up* 是一种非连续性搭配(discontinuous collocation)。非连续性搭配是界定跨距时要考虑的重要问题。肯尼迪（Kennedy 1990）发现，在 LOB 语料库中，*different from* 这个搭配共出现了 35 次，其中 14 次是以非连续性搭配的形式出现的。而在其中的 2 个非连续性搭配实例中，*different* 和 *from* 被 6 个词隔开，如：

Non-cooperators were not different in age or other conventional factors from the rest.

这就是说，与节点词相互预见，构成搭配的并不总是那些与之紧密毗邻的词；与节点词相距远的词完全可能仍与其相互吸引，构成搭配。在这一点上，词语搭配不同于复合词（compounds）；复合词是被句法关系的力量紧密地粘合在一起的一种语言单位。

在词语搭配研究实践中，跨距的界定是个颇有争议的问题。对这个问题的看法往往反映出研究者对词语搭配所采取的理论立场。限于篇幅，我们在此不作详述。但是，有一点是肯定的，即跨距是基于语料库证据的词语搭配研究必不可少而且非常有用的一种手段。多数研究项目表明，将跨距定在 −4 ／ +4 或 −5 ／ +5 是较为合适的。马丁(Martin 1983:84)等人的研究证明，如果将跨距定为 −5 ／ +5，就可以获得关于节点词 95% 多的搭配信息。琼斯和辛克莱(Jones & Singlare 1974)从 −1 ／ +1 的跨距开始，一直尝试到 −10 ／ +10 的跨距，检查在不同的跨距内所提取的搭配词的情况。结果发现，超出 +4/−4 的位置后，节点词已不再有太大的吸引力。在 −4 ／ +4 的跨距内，搭配词的分布与语法结构紧密相关；一定语法类别的词一般都如期出现在相应位置上。比如，*good* 一词的副词搭配词(adverb collocates)一般都出现在 N − 1 位置上：*jolly*

good、*extremely good*、*pretty good* 等；而介词搭配词(preposition collocates)和名词搭配词(noun collocates)一般都出现在 N + 1 位置上：*good for*、*good at*、*good fun*、*good thing*、*good heavens* 等。所以，在辛克莱及其 COBUILD 同行所进行的研究中，跨距一般定为 -4 / +4，另外一些研究者或界定为 -4 / +4，或界定为 -5 / +5。

然而，跨距的界定一定要采取灵活的策略，要视所研究文本的具体性质和特点而定。文本的语体(language variety)、文体(style)和语域等因素对跨距的界定有直接关系。有的文本，如学术英语文本的平均句长及其所用的句法结构、词语表达方式等与普通英语文本都有所不同，这就会对相互搭配的词项间的距离产生影响。所以，在具体的词语搭配研究项目中，一般应进行先期研究(pilot study)，看看跨距定为多大才能将绝大多数的搭配词包括进去，然后再作出决策。

2.3 统计检验

词语搭配研究的是词项的典型共现行为。典型性从根本上来说是个概率问题。在语法描述中，人们可以以对比的方式和较为绝对的措词来描述语法类别的选择：叙述过去时间里发生的事或动作时，要用一般过去时，而不用一般现在时，如 *He studied the collocations in EAP texts last semester*，而不是 *He studies the collocation in EAP texts last semester*。表达某事"业已完成"，要用完成体而不是进行体，如 *He has studied the collocations in EAP texts*，而不是 *He is studying the collocations in EAP texts*。而在词语行为研究中，研究者只能用概率的方式来描述搭配的典型性程度，因为词语间的任何搭配都是可能的。比如 *This lemon is sweet* 这个搭配虽然不同寻常(unusual)，但如果它是两个妇女在谈论所织坐垫套上的图案时发出的评论，或者是对一个孩子作的柠檬画发出的感叹，那就完全可以理解了(McIntosh 1967:188)。所以辛克

来说,"事实上没有不可能的词语搭配,但是一些搭配比另一些更为恰当"("There are virtually no impossible collocations, but some are more likely than others")(1966:411)。那么,就有必要对共现的词语序列进行统计检验,看其是否为显著搭配。照马丁的说法,"显著搭配是指那些其构成成分共现的次数比人们根据它们在文本中各自出现的频数预测的还要多的词语序列"("A significant collocation is one in which the two items co-occur more often than could be predicted on the basis of their respective frequencies in the length of text under consideration")(Martin, et al. 1983:84)。

在研究实践中,有几种统计方法用来测量词语搭配的显著性。一种是用搭配序列出现的频数除以节点词出现的频数,即:

搭配序列的频数÷节点词的频数

这样所得到的商总是小于或等于1,越接近于1,所检验的搭配出现的概率就越大(Martin 1983:86)。这种方法主要用来检验一个搭配序列在文本的不同部分,随着节点词频数的变化而可能出现的变化概率。这对分析研究文本各个部分的用语特点和风格具有一定的意义。

另外一种常用的概率检验是根据搭配词的相对频数(relative frequency)进行的。相对频数指的是一个搭配词在语料库中的总频数与其期望频数(expected frequency)的比率。一个搭配序列在库中实际出现的频数越高于其期望频数,它就越显著。杰尔默曾用 *to be* 和 *guerilla warfare* 两个搭配序列来说明这一点:"如果两个序列在语料库中出现的频数相同,那么,*to be* 就没有 *guerilla warfare* 显著,因为 *to* 和 *be* 这两个词在一般语料库中各自出现的频数要比 *guerilla* 和 *warfare* 两个词各自出现的频数高得多"("The more frequent a sequence is in relation to its expected frequency of occurrence, the more distinctive it is likely to be, other things being equal. If *to be* occurs a certain number of times in the corpus, this is less indicative of the distinctive-

ness of the collocation than, if, say, *guerrilla warfare* should occur the same number of times, because the individual words *to* and *be* are much more frequent in the corpus than are *guerrilla* and *warfare*")(1984:166)。

在这里,期望频数是个很重要的概念。如果把一个语料库包含的总词数假定为 W,而某个搭配词与节点词共现的频数为 C,2S 为限定的跨距,N 为节点词在语料库中出现的频数,那么这个搭配词与节点词共现的期望频数 E 就是:

$$E = C \times (2S+1) \times N \div W$$

期望频数被用于两类常用的统计度量方法中,一类是 Z 分值(或 T 分值),另一类是相互信息值(Mutual Information Value)。关于 Z 分值和 T 分值的计算与使用,参阅本书第四章、第五章 。在本节里,我们着重谈一下相互信息值(MI)的特点及其与 T 分值的不同之处。

T 分值的高低主要取决于节点词与搭配词共现的次数。共现的次数越多,T 分值便越高;反之,则越低。T 分值给研究者一种把握,使他可以判断共现的词语间在多大程度上存在着典型搭配关系。对 T 分值高的序列,研究者便有足够的把握确定其为搭配,因为统计数据表明共现的频数已很高,足可以排除是偶然共现。相互信息值测量的是词语间的搭配强度,即一个词的出现所能提供的关于另一个词出现概率的信息。克里尔(Clear 1993)曾用下述的例子来说明 MI 的原理:在一个 1000 万词的语料库中,*kin* 这个词出现了 10 次。这意味着 *kin* 在库中出现的概率是 0.000001。但是,在同一个语料库中,如果 *kith* 这个词出现了 5 次,而且在 5 个实例中,"kin"总是出现在 *kith* 之后(*kith and kin*)。那么,当我们看到 *kith* 时,我们就有 0.5 的概率看到 *kin*。这样,*kith* 的出现给我们提供了大量的信息,来揭示 *kin* 的出现(ibid:278)(关于 MI 值的计算方法,参阅 Church and Hanks 1990)。MI 值的差异表明词语搭配的不同的显著性。表 3.3 显示的是节点词 *knowledge* 的

15个 T 分值最高的搭配词的有关信息，数据来自 5000 万词的 COBUILD 语料库。

表 3.3 T 值统计测量数据

Collocates	Total occurrences	Co-occurrences with Node	T-score
lack	4170	57	7.244 933
common	6744	57	7.056 728
prior	1445	42	6.357 656
gained	1701	39	6.094 638
basic	4347	36	5.600 056
secure	2272	29	5.152 265
public	19 172	41	4.750 265
detailed	1648	22	4.496 458
general	15 785	35	4.443 188
gain	2507	22	4.395 361
intimate	707	19	4.269 362
academic	1885	19	4.120 176
medical	6433	23	4.055 357
acquired	1029	17	3.985 337
practical	2765	18	3.882 875

由表中数据可以看出，搭配词与节点词共现的频数是 T 分值高低的关键因素。共现的频数越高，T 分值也就越高，也就越能说明搭配词与节点词间存在着搭配关系。

在同一个 COBUILD 语料库中，与节点词 *knowledge* 共现的 MI 值最高的 15 个词及其有关信息显示在表 3.4 中：

表 3.4　MI 值统计测量数据

Collocates	Total occurrences	Co-occurrences with Node	MI Value
encyclopaedic	22	9	9.534 417
tacit	80	8	7.501 793
impart	70	6	7.279 379
firsthand	65	5	7.123 244
accumulating	80	4	6.501 693
encyclopedic	60	3	6.501 693
broaden	190	8	6.253 741
acquiring	266	11	6.227 743
requisite	104	4	6.123 144
grammatical	79	3	6.104 764
thorough	417	15	6.026 561
intuitive	149	5	5.926 324
prior	1445	42	5.719 007
intimate	707	19	5.605 893
gaps	338	9	5.592 577

将表 3.3 和表 3.4 所列数据比较一下就会发现 MI 测量方法与 T 分值测量方法的不同。*prior* 一词虽出现在了两个表中，但所处的次序位置极不相同：在 T 分值表中位于第三，在 MI 表中却排至第十三位。MI 值高的搭配词不一定和节点词共现的频数就高。起决定作用的是搭配词与节点词共现的频数与搭配词在语料库中出现的总频数的比值。MI 测量方法的缺点是，当一个搭配词和节点词共现的次数不多而仍有较高的 MI 值时，它可能表明了一般情况下词语的搭配力，也可能是由于语言使用者独特的个人用语特点或者某个语料库的特点所致。对此，研究者很难作出判断。MI 测量方法的优点在于它能较好地识别复合词、固定词组、科技术语等等。

究竟是使用 T 分值，还是使用 MI 值，要视研究的内容与目的

而定。

2.4 偶然共现词的排除

在上述几节中，被称为共现词的词项是所有那些落入界定跨距内的词项。这些词项可能和节点词具有语法和语义关系，也可能和节点词没有任何语法和语义关系。比如2.1中用过的那个例句中，*have*、*in*、*various* 和 *ways* 等词和节点词 *define* 并没有明显的语义和语法限制关系，但由于它们位于界定的 −4 / +4 的跨距内，也被视为 *define* 的共现词。这类共现词在研究中被称为偶然共现词。所谓偶然共现词，是指那些对节点词没有预见作用但又落入界定跨距内的词项。严格说来，偶然共现词对真正意义上的词语搭配研究没有太大意义，应当排除。排除的办法之一就是通过统计检验。一般来说，这类共现词重现的概率较低。当一个词在语料库中出现的频数较高时，统计检验会突出它的典型搭配；它的偶然共现不太可能具有统计意义上的显著性而被过滤掉。另外一个办法是对统计检验设置起点频数(threshold frequency)：搭配词只有和节点词共现达到一定频数时，才可进行统计检验。在辛克莱的研究中，起点频数定为3，因为他发现只出现一次的搭配序列可能是语言使用中的偶然行为，而与节点词共现两次的词大多也都是偶然共现词(Jones & Sinclair 1974)。一般来说，偶然共现词排除或过滤掉后，剩下的就是对节点词具有预见作用的搭配词，这些搭配词如果达到统计上的显著性标准，就和节点词构成显著搭配。这些显著搭配基本上反映了典型的词语行为，可以据此勾画出语言的词语结构。比如，与"sun"经常搭配的词可能有 *bright*、*hot*、*shine*、*light*、*lie*、*come out* 等，与 *moon* 经常搭配的词可能有 *bright*、*full*、*new*、*light*、*night*、*shine* 等。这两组词分别组成"sun"和"moon"各自的搭配范围(collocational range)(McIntosh 1967: 189)，或者叫它们的词语集(lexical set)(Halliday 1976: 80)或者叫词语型式(lexical patterning)(Sinclair 1966:411)。

2.5 定性分析与定量分析相结合的研究

上面讨论的研究方法是基于语料库证据的词语搭配研究的基本方法。它所包括的一系列技术手段和环节都是为了借助于计算机更快地、更有效地对词语行为进行检索、统计和分析。这种研究方法，就其主要本质特征来说，是一种定量分析的方法。定量研究是一切实证主义科学研究所采取的基本方法，然而，仅有定量研究是不够的。在很多情况下，加上定性研究，效果便会好得多。就词语搭配研究而言，所谓定性研究，就是根据词语搭配的主要界定特点，如因循性、意义—形式限制性、语域限制性等等，对共现的词语进行分析和描述。比如，处理偶然搭配词时，一方面借助统计手段，我们可以将它们过滤掉，另一方面我们可以根据词语搭配的意义—形式限制性这一界定特点将它们排除；按照意义—形式限制性这一标准，只有出现在同一语法框架内互为限制的词语才可能具有搭配关系，那么 2.4 中谈到的 *have*、*in*、*various* 和 *ways* 等不符合这一标准，研究者可以据此将它们排除在"define"的搭配词之外。

定性研究的必要性还体现在对少数非显著搭配的处理上。并非所有的非显著搭配都没有反映词语的典型行为。有些序列明显地反映了词语的典型性行为，但由于在语料库中出现的频数较少，所以在进行统计检验时便不够显著。出现这种情况的原因是多方面的。可能是由于所依据的语料库不够有代表性，也可能是语域、语体等具体因素所致。在这一点上，研究者应当清楚，任何语料库，即使其容量再大，包括的文本再全面，从根本上说来，都只是对真实语言使用的有限抽样。所以总难免漏掉一些常见的语言现象，或者导致这些常见的语言现象未能在库中充分反映。*raise army* 这一习惯性搭配在 5000 万词的 COBUILD 语料库中的情况就是如此。该语料库的容量可谓超级规模，所包括的文本涉及数十种题材、体裁，各种风格、语体，可谓全面。在以 *raise* 为节点词进行的查询研究中，数十个常见的词语搭配都达到了统计度量上的显著性标准，

所以是显著搭配，如 *raise money*、*raise funds*、*raise cash*、*raise questions*、*raise taxes*、*raise issue*、*raise awareness*、*raisestandards*、*raise rates* 等等。但 *raise army* 却不是。对此，研究者就应采用定性分析的方法，将其视为典型搭配，而不是将其排除不予考虑。

应当说，在整个研究过程中，包括提取搭配词、统计检验、设置起点频数等具体环节，都应当采用定量分析和定性分析相结合的方法。定性分析依据的是关于词语搭配的内在标准，直接涉及词语序列的性质；定量分析依据研究者设计的一套外在标准，不一定完全反映词语序列的性质。然而从另一方面讲，定性分析靠研究者个人的判断，可能带有主观意识；定量分析则完全是客观的，靠真实数据支持的。没有定性分析，词语搭配的内在性质可能会得不到充分的重视，使研究结果不尽正确；没有定量分析及其一整套研究手段和环节，研究就失去了坚实的数据基础，研究者也不可能发现大量的在真实语言交际活动中被使用的搭配序列；语言经验再丰富的研究者，其直觉都是有限的和靠不住的。所以定量研究和定性研究相结合的原则应是基于语料库证据的词语搭配研究应遵循的重要原则。

3　词语搭配研究的语言学意义

在上述几节里，我们主要是把词语搭配看做一种语言使用现象，一种语言单位(linguistic unit)来研究。然而从更高一级层次上来看，词语搭配还是一种语言研究方法(a linguistic study approach)。这种研究方法可概括为词项—语境方法(item-environment approach)(Francis 1993:146)。它不失为一种科学和有效的方法，对语言学各个层面的研究都具有深刻意义。

3.1 词语搭配研究：融词汇与语法于一体

语言运作，就其本质而言，是由词语与语法组成的一个连续体的运作。从理论上来说，语法与词语是在两个不同层面上对语言进行的抽象。前者研究的是制约语言使用的规则，限制着如何将词语组织成为结构良好的句子，后者探讨的是个体词项的具体行为。然而，问题的另一面是，词语和语法是相互渗透和相互交织一起的，很难截然分开。理想的语言研究应将二者综合一起：任何词汇行为的描述都应参照该词常用的语法结构来进行；任何语法结构的描述都应基于相关词语的真实行为而进行。而词语搭配这个语言单位正是语法与词汇的汇合点(where grammar and lexis meet)。在搭配这个层面上，我们所观察到的语言实体(linguistic entity)融词和语法为一体，集形式和意义于一身；语法结构也是词语结构，词语的行为模式也是语法结构——究竟看做什么则取决于研究者的观点和研究目的。用搭配研究中常用的行话来说，类联接与词语行为融合到了一起。下面，我们用英语的动者格动词(ergative verb)结构来说明这些问题。

动者格动词指那些既可用作及物动词又可用作不及物动词而不改变动词本身意义的词，如 *open*、*break*、*stop* 等词的用法。

When I opened the door, there was Laura.
Suddenly the door opened.
I broke the glass.
The glass broke all over the floor.
The driver stopped the car.
A big car stopped.
（上例参见 Sinclair 1991b:152－153)

显而易见，动者格动词的及物性用法和不及物性用法之间是有内在联系的：及物性用法使人知道施动者，也就是主语表示的人或

物;不及物性用法的信息焦点在动作本身而不在施动者,但不及物性用法的主语就是及物性用法的宾语。然而,作动者格动词的主语或宾语的词是有一定的范围限制的,每个动词的范围都不一样。以动词 *show* 为例。语料库证据表明,常作"show"的主语或宾语的是那些表示感情因素的名词,如 *anger*、*joy*、*disappointment*、*emotions*、*fear* 等:

He showed his anger.
His anger showed.
He showed his joy.
His joy showed.
Her face must have showed her disappointment.
Her disappointment showed.
She has showed her emotions.
Her emotions showed.

(参见 Sinclair 1990:157; 1991b:152-153)

在词语搭配研究中,研究者要参照类联接来描述 *show* 的行为,诸如"及物性用法"、"不及物性用法"、"主语"、"宾语"等语法术语。他还要用"节点词"、"搭配词"、"搭配范围"等关于词语行为的概念来描述与 *show* 搭配的一连串名词: *joy*、*disappointment*、*emotions* 等等。最后,他会做出这样的研究结论:节点词的典型搭配词包括 *joy*、*anger*、*disappointment*、*emotions*、*fear* 等词,这些词既可作 *show* 的主语,构成 N+V 搭配,又可作 *show* 的宾语,构成 V+N 搭配。在这个过程中,语法和词语的描述就综合在一起构成了有机的整体。不描述词语行为和词语搭配范畴,语言描述就不完善,很可能是对真实语言使用的曲解和对语言学习者的误导;不参照语法结构,词语搭配的解释力将是有限的。二者实在是一个硬币的两个面。

有些动者格动词用作不及物动词时,往往要用一个副词做状

语，比如 *sell*；在动者格结构中，如果 *sell* 用作不及物动词，它大都要求 *well* 这个副词作状语。下面是从 COBUILD 语料库中提取的词语索引：

1	The Squadron range is starting to sell	well in the States [p] The
2	[p] The parrots bred in Yunmeg now sell	well in the country's coastal areas.
3	in 1949, but was too unorthodox to sell	well-unlike the Standard Eight,
4	Jaguar SJ12 Coupe. A car which did sell	well was the 1934 Standard 'Hot Rod'
5	books in the Anna Pigeon series have sold	well here and abroad, finding their
6	as had Dwight, that the book had not sold	well and that some Polish e *
7	year in that desert state. Her work sold	well enough to make her comfortably
8	in the history of sport that had sold	well under a star's imprimatur: The
9	the car with its Chrysler V8 engine, sold	well until the 1973 fuel crisis and
10	Sotheby's Hendon auction on July 10 sold	well, while the cheaper cars at
11	Penman-bodied Rolls-Royce 20hp, which sold	well. Two less-exalted but
12	cars the 1935 Austin Hereford saloon sold	well. [p] Looking to the future
13	and four volumes on pedagogy, which sold	well enough to encourage him to try
14	Headline, the publisher, it has sold	well since its publication last year
15	American and Latin American art all sold	well in the spring. Mujer Con
16	popular. Progress and Poverty sold	well over a million copies. The

由上面的索引可以看出：在不及物用法中，动者格动词 *sell* (*sold*)的主语一般是指 *book*、*car*、*pet animal*、*work of art* 等类事物。另外一个明显特点是，*sell* 要类联接于一个副词性状语，但它几乎无一例外地与 *well* 搭配，构成 V + ADV 搭配。

同样，在这样的词语搭配描述中，语法规则与词语行为融合一起，相互参照，相互交织。这种词项—语境研究方法是词语搭配研究的正确方法，也是一般语言学描述应当采取的适当方法。在语言使用中，交际者对语言形式的选择绝不是"先选择语法结构，然后用词语来填充结构"。语言学描述中的空位填缺(slot-filler)的模式是脱离语言使用实际的模式，应当改进，而词语搭配研究无疑给语

言学家提供了有意义的方法。

3.2　词语搭配研究：客观、具体的典型意义研究

在前面，我们说过，词语搭配是“意义—形式的综合体”，具有“意义—形式限制性”的界定特点。这意味着，对意义的研究是词语搭配研究的重要内容。然而，不同于传统的语义学研究，词语搭配研究的是客观的、具体的典型意义。它对意义的研究对一般语言学中的意义研究具有一定的启示意义。

词语搭配研究的意义是语言形式在语境中所实现的具体意义。对意义的研究从语言内部的运作开始，以真实使用的语言材料为对象，而不是从语言外部着手，照外部世界的模式对意义进行切分、抽象。外部世界一些事物的结合模式当然在词语搭配的模式中有所反映，如 *arms and legs*、*desks and chairs*、*bread and butter*、*blood and flesh* 等等。但是，词语搭配研究并不从这些有限的反映外部世界事物结合模式的词语组合或搭配出发，挖掘或抽象它们可能具有的意义体系。强大的语料库证据及其有效的研究方法和手段使词语使用的典型语境和形式都呈现在研究者面前，研究者则可以依据大量的真实搭配实例，对词语所实现的典型意义进行分析、概括和归纳。因此，词语搭配研究的焦点是在语言交际中被经常使用的普遍的、习惯性的搭配形式及其意义。这些是客观的意义、具体的意义和典型的意义。下面，我们以 *career* 一词的搭配行为及其实现的意义为例来说明。从 5000 万词的 COBUILD 语料库的证据来看，*career* 一词的搭配范围包括数十个词，其中最常用的习惯性搭配词及其有关信息如下表所示：

表 3.5 *career* 一词的左搭配词数据统计

Collocates	Total Occurrences	Co-occurrences With node	T-score
international	14 167	145	10.857 743
political	14 288	116	9.435 439
professional	5689	96	9.213 702
successful	5113	90	8.944 510
started	12 002	92	8.332 555
start	16 476	94	7.985 379
pursue	913	63	7.821 509
playing	9873	71	7.247 124
glittering	377	53	7.228 002
acting	2721	57	7.187 179
development	9793	67	6.981 477
stage	9257	66	6.977 466
distinguished	811	48	6.810 415
football	17 572	72	6.401 471
managerial	417	33	5.671 519
goals	5103	41	5.601 193
racing	3603	36	5.395 751
singing	2043	33	5.38 6701
management	6358	37	5.030 987
promising	1029	27	4.996 885
training	8415	40	4.985 719
resume	701	25	4.858 925
brilliant	3378	30	4.856 639
business	17 692	51	4.648 584
academic	1885	25	4.620 646
remarkable	1834	24	4.522 278
Hollywood	2496	25	4.497 683
teaching	3035	25	4.389 210
writing	5570	29	4.344 384
musical	2218	23	4.330 459

仔细观察，不难发现，表中的搭配词，无论是名词、动词，还是形容词，都有一个共同的语义特点，即表达一种积极的内涵意义。这是和“career”有关的词语所实现的意义的一大特点。*career* 的名词搭配词和形容词搭配词大致可分为两类，一类是评述性的搭配词，另一类是描述性的搭配词。现以形容词性搭配词为例，举例如下：

评述性搭配词：*international*，*successful*，*glittering*，*distinguished*，*promising*，*brilliant*，*remarkable*.

这些评述性的搭配词都具有积极的内涵意义，无一是表示消极意义的词；*international career* 显然是令人羡慕的职业，*glittering career*、*distinguished career*、*promising career* 等等则表达语言使用者对相关职业的高度赞许和肯定。

描述性搭配词：*political*，*professional*，*playing*，*acting*，*managerial*，*racing*，*singing*，*academic*，*musical*.

这些描述性搭配词说明 *career* 的种类或所属的行业领域；无论是 *political career*、*professional career*、*managerial career*、*academic career* 还是 *musical career*，在英国社会都是高度受人尊重，或具有很高社会地位的职业；而 *racing career*、*singing career* 和 *acting career* 则是受人瞩目或令人羡慕的职业。

所以，“career”的上述搭配词几乎无一例外地属于表达“好事物”的词语，对“career”赞许、肯定并传递一种愉快的感情。它们似乎在“career”的语境中形成一种强烈的积极语义氛围；所有的搭配词之间都似乎存在着一种语义和谐；任何不和谐的、和积极的语义氛围相冲突的词语都会被排斥于外。而动词搭配词则主要表示一些具体的动作意义：*career* 可以“开始”（*started*），可以被人们“追求”和“从事”（*pursue*），可以“恢复”（*resume*），可以“管理”（*managed*）等等。

这些就是和 *career* 有关的词语搭配所实现的典型意义及其具有的语义特色。它们是十分客观和具体的意义。

词语搭配研究对意义的探讨揭示了一种更为直接的和切合实际的研究方法，它至少是对传统语义学研究方法的一种补充。那些

撇开真实的语言使用证据，对意义进行的切分、抽象，会产生深奥的理论，但这些深奥的理论究竟在多大程度上反映了语言使用中的实现意义(realized meaning)和典型意义，尚有待于检验。

3.3 词语搭配研究：对孕于形式中的功能进行研究

所谓功能，就是指用语言做事。语言的意义和功能也处于一个连续体内，只是做事的程度不同而已。词语搭配也可以说是融形式、意义和功能于一体的语言单位。就语篇功能而言，所有的词语不同程度上都起着对篇章信息组织的作用。而词语和词语搭配序列的语篇功能也是搭配研究的重要内容之一。

在前面的1.2.1一节中，我们曾谈到 *This paper describes* 这一词语搭配序列。事实上，这个搭配序列在学术论文中的语篇功能是十分明显的。一般的学术论文都有引语(introduction)一章。引语一般又由三个语步(move)组成。而最后一个语步“占领位置”(Occupying the niche) 大致上包括 *Outlining purposes*、*Announcing present research* 等语阶(step)(见 Swales 1990:141)。我们所谈的 *This paper describes* 这个搭配序列正是在语篇中起着 *Announcing present research* 这个功能，即向读者宣布论文所报道的主要研究内容。除这个序列外，被学术研究人员用来“宣布论文内容”的搭配序列还有

This paper presents ...

This paper discusses ...

This paper examines ...

This paper investigates ...

This paper reports ...

This paper deals with ...

等等(参见本书第六章)。

研究者在研究这些搭配序列时，不可避免地要对它们的语篇功能进行研究。研究的特点是通过观察搭配序列的形式构成，检查搭

配序列所处的语境，前后语境中的语篇信息特点，参照整个语篇的结构，从而确定搭配序列实现的语篇意义和功能。在研究者看来，形式、意义和功能都孕育在搭配序列中，三者是不可分割的一体(oneness)，所以形式、意义和功能紧密结合起来研究是词语搭配研究的重要特点。

结 束 语

本章从一些基本的背景知识和重要语言学观点谈起，讨论了词语搭配研究的主要界定特点和基本框架体系，描述了基于语料库证据的词语搭配研究的基本方法和路径及其具体环节，并探讨了词语搭配研究对一般语言学研究产生和可能产生的意义和启示。总的来说，词语搭配是语料库语言学实践中最富活力的研究活动之一，它揭示了一种新的、科学的研究语言的思想、方法和原则。

在目前进行的以自然发生数据支持的研究活动中，词语搭配这个学术界久已谈论的话题被赋予了全新的内容、含义、理念、思想和方法。本章对这些内容进行了概述，对含义进行了探讨，对理念进行了界定，对思想和方法进行了介绍。从另一方面讲，词语搭配研究仍在向前发展，新的研究内容会开辟出，更科学的思想和方法也会问世并可能取代现有的思想和方法。我们应不断研究、总结和吸收新的思想和方法，丰富词语搭配研究的框架体系，使之更科学、更富有生命力。



第四章　语料库建设及其基本统计手段和原理

引　言

语料库语言学正以其独特的优势迅猛发展。语料库已被广泛应用于与语言相关的各个领域,如语言学研究、语言教学与研究、自然语言处理和语言工程等,并在其中发挥日益重要的作用。因此,语料库建设也就显得越发重要。语料库的产生使定性与定量相结合的语言研究成为可能。著名语言学家韩礼德(M. Halliday)明确指出,"语言系统天生就是概率性的"("the linguistic system was inherently probabilistic")(1991:31)。这就是说,语言中的概率信息是语言本质重要的组成部分。如何利用必要的统计手段把这些概率信息提炼出来,也是语料库语言学的重要任务之一。语料库语言学家正从各个角度对语言进行分析处理以便挖掘出有意义的概率信息。本章首先讨论语料库建设中需要特别考虑和解决的几个问题,即语料库的代表性问题和预加工问题。然后介绍当前用于语料分析的几个主要的统计手段及其原理。

1　语料库建设

随着计算机及其相关技术的飞速发展,语料库建设变得越来越容易。扫描仪、电子出版物以及网上资源为收集语料提供了极大的

便利。建立语料库已经不再是专家们的专利,广大的语言研究工作者已经完全可以自行建立起中小型语料库,进行语言学和应用语言学研究。尽管如此,语料库建设中需要考虑和解决的问题还有不少,要建立起规模大、质量高、用途广的语料库也并非易事。语料库的代表性和预加工就是两个重要问题。

1.1 语料库的代表性问题

早在20世纪30年代计算机尚未问世之时,一些语言学家已经开始建立一定规模的非机读语料库并开始进行人工词频统计。这种基于语料的经验主义研究方法也体现在当时的美国结构主义语言学研究之中(参见 Fries 1952)。这种研究方法以其实证性而赢得了语言学家的广泛赞誉,并在50年代得到了蓬勃发展。韦斯特(M. West 1953)在非机读语料库基础上人工统计出的通用词汇表(General Service List)至今还被教材编写者参照(参阅 Willis 1990)。然而,这种良好的发展势头在50年代末受到了重创。乔姆斯基(1957)《句法结构》的发表标志着以布隆菲尔德(L. Bloomfield)为代表的结构主义时代的结束,也标志着转换生成语法的开始。在这部著作中,乔姆斯基批评了语料库研究方法。根据麦克内里和威尔逊(T. McEnery & A. Wilson)的概括(1996:4-10),乔姆斯基对语料库的批评归纳起来主要有三点:1)语料库模拟的是语言的使用(performance)而不是能力(competence);2)语料库试图列举无限的自然语言;3)语料库往往完全避开内省(introspection)。从现在的角度看,这三点批评已经不能完全站得住脚了,但是在当时却使语料库研究方法受到了严重打击,语料库语言学的发展也因此受到影响。20世纪60年代和70年代是乔姆斯基语言学的鼎盛时期,在某种意义上可以说是一统天下,自然也是语言学的绝对主流。语料库方法与乔姆斯基语言学相比而言当然只是一个微不足道的支流,但它毕竟没有断流。也正是在这期间,三个著名的机读语料库—— BROWN、LLC、LOB 相继问世,为语料库语言学的发展打下了坚实基础。从80年代开始,基于语料库

的语言学研究得到了迅速发展。到了 90 年代，大型的机读语料库如 BNC(British National Corpus)、Bank of English 等得以建立并投入实际使用，吸引了大批的语言学家投身于语料库语言学研究。正如一些语言学家指出的那样，语料库语言学正变得愈益重要(参阅 Svartvik 1996: 3)。

有人把乔姆斯基出现之前的语料库研究称为早期语料库语言学(参阅 McEnery & Wilson 1996)。乔姆斯基对早期语料库语言学所作的上述三点批评不能说全无道理，但是没有哪一点批评是真正致命的。就第一点批评而言，乔姆斯基认为语料库因为模拟的是语言的使用而非能力，因此不是研究语言的最好材料。他的这一观点是建立在对能力和使用的区分之上的。但是，有些语言学家，尤其是伦敦学派的语言学家，根本就不赞成能力和使用的这种区分。他们认为能力和使用不是两个事物，是不能截然分开的(参阅 Firth 1957, Halliday 1991, Sinclair 1991 等)。所以语料库反映语言使用不是什么不正常的事，恰恰相反正是它的长处所在，它反映的就是人们实际使用中的真实语言。至于乔姆斯基的第二点批评，他要说的实际上是语料库无法列举无限的语言。但是，对无限的总体进行抽样调查早已证明是行之有效的科学方法。语料库是无限语言的样本，用它研究语言是无可厚非的。当然，在抽样的过程中的确需要考虑样本的代表性问题。至于乔姆斯基的第三点批评，当前多数语言学家并没有否认内省的价值和意义，只是认为内省的证据需要证实。

所以，从现在的角度看，这三点批评中需要考虑的是第二点，它向语料库语言学家提出了如何解决语料库代表性的问题。语料库是否具有代表性直接关系到在语料库基础上所作出的研究及其结论的可靠性和普遍性。这里需要考虑三个问题，即语料库代表的总体、语料库的规模及语料库的内容。

1.1.1 语料库代表的总体

任何语料库的建立都有一定的目的。建库的目的与语料库的

代表性是密切相关的。一个语料库是否有代表性首先要看该语料库所代表的总体。作为一般语言资源的语料库代表的总体是无限的,或者是非常大的。就专门语料库而言,情况有所不同。比如:为了研究莎士比亚的作品,人们也可以专门建立一个语料库,收集莎士比亚的全部作品。这时候,语料库就不涉及代表性的问题,因为要研究的有限总体可以全部收集到库中。但是,如果要研究英国文学,那么莎士比亚的全部作品也不过是英国文学的一小部分,单靠这个语料库显然是不够的。事实上,在多数情况下,语料库代表的往往是无限的总体。60 年代建立的第一个机读语料库 BROWN 语料库是一个约 100 万词的旨在研究一般美国英语的语料库。由于它要代表一般英语,语料库的采样涉及的面便相当广,有 15 个领域的文本共 500 篇,每篇取 2000 词或稍多(详见 Leech 1987:7,桂诗春、宁春岩 1997:140)。1978 年完成的与 BROWN 对应的英国英语语料库 LOB 也是约 100 万词的语料库。以现在的标准来衡量,100 万词的语料库已经太小了,不管它涉及的面有多广,其代表性也是很有限的。然而在当时建立这样规模的语料库是相当不容易的,完全可以说是一项大工程。尽管从现在看来,它们的代表性还相当不够,但它们给研究者提供关于语言使用的信息已经相当可观,足以对以直觉证据为主的语言学研究构成挑战。

1.1.2 语料库的规模

如果语料库要代表一个无限的或者非常大的总体,那么就有一个采样或抽样的问题。对于一个无限的总体来说,在其他条件相同的情况下,样本越大则代表性越好。按现在的技术,建立大型的语料库已经不是一件难事。由英国伯明翰大学和柯林斯出版社联合建立的英语文库(Bank of English)是一个不固定大小的语料库,早在 1996 年已达 3.2 亿词(参阅 Rundell 1996),而且还在以每月 500 万词的速度增长(参阅 桂诗春 1997:139)。这个语料库与 BROWN 和 LOB 不同的是:第一,它是一个动态和开放的语料库;第二,它没有采取分层随机抽样的方法,即没有事先设定收集哪些

领域的文本,更没有各领域的比例。由于这些特点,有人将其称作语料档案(archive),而只将采取分层随机抽样、按照比例收集文本的LOB、BROWN 等称为语料库(corpus)(参见 Leech 1991, Kennedy 1998)。先不论这种区分是否必要,我们认为,规模自然相当重要(参阅 Leech 1991, Rundell 1996, Sinclair 1997),但是语料库中的内容也需要考虑,对文本作一些分类和归纳是必要的。

1.1.3 语料库的内容

语料库的规模和内容是一个问题的两个方面,前者是量的问题,后者是质的问题。规模大固然能弥补内容上的偏颇,但是规模大并不意味着对内容就无须作要求了。对于内容,最根本的要求是:真实。这里的"真实"有两层意思:其一是要收集实际使用中的文本(不能是语言学家或研究者自己杜撰的文本),其二是要收集符合条件的文本。如:如果要收集的是正式出版物上的文本,那么就不能收集非正式出版物或尚未正式出版的文本。又如:如果要建立学习者语料库,要研究学生真实的语言能力,就不能把学生抄袭的作文收进语料库,即使不留心收进去了,也要进行删除。一句话,在语料库中所收集的文本应该是总体的真实样本。但是,真实并非是代表性的惟一标准。

除了真实以外,还要考虑收入语料库各类型文本的比例。比如,要建立一般英语的语料库,需要考虑口语和书面语的比例。再说,还有介于口语和书面语之间的语体(如播音稿、演讲稿等)也需要考虑。BROWN 和 LOB 语料库旨在研究一般英语,但它们只考虑书面语。这些语料库的建设者们把一般英语先分成 15 个领域,然后对每个领域按一定的比例收集文本。关于如何来划分领域和确定各部分之间的比例关系,辛克莱(1997)认为,对于文本的分类,目前不宜从内部着手,因为我们对文本的内部特征的认识还相当有限,而更应从一些外部的社会标准来分类。在目前,这应是一个现实可行的办法。至于各部分之间的比例安排问题,最好也能找到一些较为客观的外部标准,如在确定分类的基础上,最好能按各类文

本平均每天出版的词数来确定各自的比例。这是应当尽量参照的客观标准。总之,不应该单凭主观的判断来决定各部分的比例。

此外,BROWN 和 LOB 语料库中的文本一般都是不完整的,因为所有的文本只收 2000 个词或稍多(都在第 2000 个词所在的句子为止)。现在看来,这种做法有利也有弊,每篇文本都是 2000 词左右,文本比较整齐,便于比较。但是,一篇文本是一个完整的单位,文本的首尾特征一般都有差异,舍弃文本后面部分显然不是很科学的做法。至少,不完整的文本不能很好地用于语篇研究。收集完整的文本也是对语料内容的一种要求,虽然适当取舍也未尝不可。

1.1.4 建议

语料库的蓬勃发展始于 20 世纪 80 年代。到目前为止,语料库建设和研究积累了许多宝贵的做法和经验。建库时内容和规模应该兼顾。在规模上,只要条件许可,应该是使语料库有最大的规模,愈大愈好。在内容上,除了采集真实语言,还要进行必要的文本分类,并按相对客观的标准确定各类的比例。

就目前的情况来看,建立固定规模的语料库不是语料库语言学发展的大趋势。正如语言本身是动态发展的一样,语料库也应该是动态的,可以不断扩充的。比如说,要建立一般英语语料库,可以像建立 BROWN 或 LOB 时那样首先确定一些大类。在大类的基础上还可以进一步分出一些次类。然后收集符合要求的文本并把它们归入各自的类别中,先不考虑各部分的比例,文本的数量自然是多多益善。等各部分都有一定规模时,该语料库就能用于一定目的的研究了。在研究的同时,这样的收集工作可以不断进行下去。可以把整个语料库看成是一个总库,总库下面可以有子库,子库下面还可以有子库。假设按 BROWN 和 LOB 中的 15 类语体去收集语料,每一类都可以单独构成一个专业子语料库,在每一类下面还可以分出许多子类,每个子类可以构成更专业的子语料库。如果有比例标准可以遵循的话,当需要研究一般英语时,可以从各类中抽出

相应比例的语料构成一个用于研究一般英语的语料库。当然，文本的分类不应局限于语体，比如出版时间、作者的国别等也可以作为分类的标准，用于历时的比较、各种英语变体的比较等。总之，语料库应能够灵活地应用于各种目的的语言研究，而不是只能作固定的、有限的使用。为了达到这个目的，需要对输入的每一篇语料进行必要的标识。例如：文本的作者、作者的国籍/母语、发表的时间、文本的标题、文本的类型、语体等等。关于标识，在下面一节中还有更详细的介绍。简言之，每一篇文本都应是独立标识的，以便需要时可以按各种不同标准进行归类，在巨大的语料库中抽出需要的文本组成临时的子语料库进行分析研究。能否实现这种灵活利用的关键是编写能识别各个分类标识并能按需要提取文本的软件。

1.2　语料库的预加工

文本输入计算机之后，一般需要进行一些预加工，主要包括语料的标识和语料的赋码(注：赋码也可以看做是一种特殊的标识)。

1.2.1　语料库的标识(annotation)

语料库的标识主要分两类：一类是对文本的性质和特征进行标识，另一类是对文本中符号、格式等进行标识。各个语料库在标识上的做法不尽相同。例如，BROWN 在对语料进行标识时，以行为单位，对每一行的出处进行标识，其他的标识非常少，可能仅限于一些特殊字体，如斜体等。相对而言，LOB 语料库的标识要详细得多。它除了对每一行进行出处的标识以外，还进行了标题、段落、句子的开始、各种字体(包括大小写、斜体等)、某些标点符号的标识，等等。当前语料库很多，标识各不相同，在它们基础上编写的软件也不能通用，这给语料库语言学家之间的交流和协作带来了诸多困难，因此许多语料库语言学家呼吁统一标识。为此，还建立了一个专门机构来统一标识符，以便于各语料库及软件的共享。也有一部分语料库语言学家强调无标识码的原文本(clean text)的重要性：

无标识码(主要指没有关于字体、段落、句子等的标识码)的文本是非常有用的。一般的索引软件更适用于无标识的语料。总的来说,不管语料将来作什么用,第一类标识是必要的,因为它们可以用来对文本进行必要的分类,为灵活提取文本进行各类目的的研究提供了极大的便利,而且它们可以标注在文章的开头或者作为另一个文件保存,丝毫不破坏语料的完整性和原始性;至于第二类标识,对于某些研究和应用(如自然语言处理)也是必要的,但不管怎样,保留未标识之前的原文本也是必要的。

1.2.2 语料库的赋码

上面已经提到,许多研究不需要赋码语料库。但是,赋码语料库有其重要价值。当前,语料库的赋码主要是两类:一类是词类码,又称语法码;另一类是句法码,一般称为句法分析(syntactic parsing)。这两类码都为进一步的研究和应用服务。赋码的方法有多种,这里主要介绍兰开斯特大学的研究小组 UCREL(Unit for Computer Research on the English Language)的赋码方法,目的是使尚不了解赋码原理的读者对赋码有一个基本认识(关于其他词类赋码和句法分析方法可参阅:陈建生 1998;冯志伟 1997)。

1.2.2.1 词类赋码(tagging)

所谓词类赋码就是对文本中每一个单词赋予相应的词类码,包括对标点符号的赋码。词类码代表了一个词的语法特征,所以也称作语法码。这里所说的词类码,通常在传统语法词类划分的基础上进行细分,以传达尽可能多的语法信息。例如:在 LOB 中,NN 代表普通名词的单数形式(singular common noun),如:*boy*、*desk*;NNP 代表以大写字母开头的普通名词的单数形式(singular common noun with word-initial capital),如:*Englishman*、*Londoner*;NNS 代表普通名词的复数形式(plural common noun),如:*boys*、*desks*。就这样,名词被细分成几十类。动词也一样被细细地作了区分。例如:VB 代表动词的基形(base form),如:*eat*、*request*;VBD 表示动词的过去式(past tense),如:*ate*、*requested*;VBG 表示动词的现在分词形

式(present participle),如:eating、*requesting*;VBN 表示动词的过去分词形式(past participle),如:*eaten*、*requested*;VBZ 表示动词现在时第三人称单数(3rd person singular form),如:*eats*、*requests*。在对 LOB 的词类赋码中一共使用了 133 个不同的词类码(见附录,或 Sampson 1987: 165－183)。

如果英语单词用法单一,一个单词只属一种词类,那么词类赋码将会非常简单,然而事情并非如此简单。英语单词,尤其是常用词的用法一般比较复杂,一词多义的现象相当普遍。一词多义不仅体现在一个词可以用作不同的词类,而且还体现在用作同一词类时它还可以传达不同的意义。在赋码操作中,需要解决的问题自然是区分一个词不同的词类,而对具体的词义不作区分,因为这不是赋词类码的主要目的。这里所讲的词类赋码是计算机自动赋码。语料库建设者可开发一系列软件用于识别和区分不同词类。我们以 UCREL 研究小组使用的 CLAWS 系统为例来说明如何让计算机区分词类。

首先,要编出一本标明单词词类的词典。这本词典要标出一个词所有的词类,对各个词类要进行一定的排列。排列的方法自然应该是按使用频率。例如,*round* 一词在词典中是这样描述的:

round　JJ RI NN VB@ IN@
(JJ、RI、NN、VB、IN 分别代表:一般形容词,与介词同形的副词,单数普通名词,基形动词、介词;@代表该词类使用概率很低)

单单以频率排列的顺序去赋码,错误率显然相当高,必须充分考虑各个可能的词类才能提高赋码的准确率。为此,UCREL 的研究者应用马尔可夫第一过程,对 133 个词类码作了一个概率矩阵。该矩阵标明了各个词类码跟在某一词类码后面的概率。为了方便说明问题,假设我们所用的词类码只有八个,如:N、V、J、R、P、C、A、I,分别代表名词、动词、形容词、副词、介词、连词、冠词和叹词。概率矩阵必须提供 N 后跟 N、V、J、R、P、C、A、I 各自的概率,

V 后跟 N、V、J、R、P、C、A、I 各自的概率,J 后跟 N、V、J、R、P、C、A、I 各自的概率,等等。用表格来表示就是一个矩阵:

表 4.1 词类码概率矩阵

	N	V	J	R	P	C	A	I
N	x	x	x	x	x	x	x	x
V	12	13	10	17	21	8	16	3
J	x	x	x	x	x	x	x	x
R	x	x	x	x	x	x	x	x
P	x	x	x	x	x	x	x	x
C	x	x	x	x	x	x	x	x
A	x	x	x	x	x	x	x	x
I	x	x	x	x	x	x	x	x

表中列出了 8 个词类后跟各词类码的概率。如 V 后跟 N 的概率是 12%,后跟 V 的概率是 13%,后跟 J 的概率是 10%,后跟 R 的概率是 17%,后跟 P 的概率是 21%,后跟 C 的概率是 16%,后跟 A 的概率是 8%,后跟 I 的概率是 3%等等。不管是词典还是概率矩阵都要依赖于一个已赋好码的语料库。没有赋好码的语料库就无法知道当一个词有一个以上词类时后跟某一词类的概率是多少。通常的做法是先对小规模语料用手工赋码,求出初步的词类相邻码渡越概率信息矩阵,用于处理规模大一些的语料,根据处理结果,修正词类相邻码渡越概率,如此循环,直到得到稳定的词类相邻码渡越概率信息为止。若已有成熟的词类相邻码渡越概率信息,则当然可以借用,例如在对 LOB 赋码时,这些信息都来自于已赋码的 BROWN 语料库。

有了以上的词典和概率矩阵后,下一步就可以着手解决一词多码的问题了。现在以 CLAWS 对 *Henry likes stews*. 这个句子的赋码为例来说明它是怎么解决一词多码问题的。首先,把此句中的

每一个单词(包括标点)依次到事先编好的词典中去查语法码。如果某个单词有一个以上的语法码,就按出现频率的大小依次附在单词之后。这样一来,上句的赋码结果如下:

Henry	NP	
likes	NNS	VBZ
stews	NNS	VBZ
.	.	

NP 代表单数专有名词(singular proper noun);NNS 代表复数普通名词(plural common noun);VBZ 代表动词第三人称单数(3^{rd} person singular form of verb);"."代表句号本身(full stop)。

从这个结果可以看到,只有 Henry 一词的词性是确定的,作为专有名词,只有一个语法码,而 *likes* 和 *stews* 则不同,它们既可以是名词,也可以是动词。如果不考虑概率信息,这句话的赋码结果就有四种可能,即:

NP-NNS-NNS-.
NP-NNS-VBZ-.
NP-VBZ-NNS-.
NP-VBZ-VBZ-.

这就产生了歧义,但是,一个句子一般情况下只有一种解释。究竟哪一个序列最有可能,要根据计算的结果看哪一个序列出现的概率大。序列概率的计算方法如下:
先计算一个语法码后面跟另一个语法码的渡越概率,其公式如下:

$$\frac{\text{语法码 } A \text{ 与 } B \text{ 在语料库中的频数}}{\text{语法码 } A \text{ 在语料库中的频数}} \times 1000$$

按照这个公式，可以计算出 NP 后跟 NNS 和跟 VBZ 各自的概率，NNS 后跟 NNS 和 VBZ 各自的概率，VBZ 后跟 NNS 和 VBZ 各自的概率，以及 NNS 和 VBZ 后跟句号“.”各自的概率。根据 BROWN 语料库，它们各自的概率如下：

NP-NNS	17%	NP-VBZ	7%
NNS-NNS	5%	NNS-VBZ	1%
VBZ-NNS	28%	VBZ-VBZ	0%
NNS-.	135%	VBZ-.	37%

上表第一行表示，NP 后跟 NNS 和 VBZ 的概率分别是 17%和 7%，其他同理。有了以上这些概率信息，就可以计算出四个序列各自的可能性值，方法如下：

val(NP-NNS-NNS-.) = 17 * 5 * 135 = 11 475
val(NP-NNS-VBZ-.) = 17 * 1 * 37 = 629
val(NP-VBZ-NNS-.) = 7 * 28 * 135 = 26 460
val(NP-VBZ-VBZ-.) = 7 * 0 * 37 = 0

如上所示，第三个序列的可能性最大。最后经过下面公式的计算就可以求出每一序列的概率：

$$prob(X_i) = \frac{Val(X_i)}{\sum_{i=1}^{n} val(X_j)}$$

其中，第三个序列的概率是：

$$\frac{val(NP - VBZ - NNS - .)}{val(NP - NNS - NNS - .) + val(NP - NNS - VBZ - .) + \ldots} =$$

$$\frac{26\ 460}{11\ 475 + 629 + 26\ 460 + 0} = 69\%$$

这也就是说,第三个序列的可能性有百分之六十九,而其他序列的可能性就小得多,其中最后一个序列的可能性为零。至此,计算机可以初步断定这一句的赋码结果应是:

Henry	NP
likes	VBZ
stews	NNS

以上是 CLAWS 系统解决一词多码问题的基本方法。严格地讲,CLAWS 赋码系统分五个步骤对文本进行赋码。它们依次是:a)预编辑(pre-editing);b)配码(tag assignment);c)习语赋码(idiom-tagging);d)解决歧义码(tag disambiguation);e)后编辑(post-editing),其中 b)、c)、d)是赋词类码的核心步骤(参见 Garside 1987)。预编辑是由一个叫做 PREEDIT 的程序来完成的,其主要目的是为语料库中每一个词或标点创建单独的一行,把词和标点放在这一行中规定的标准位置,并用字母和数字标出该词或标点在语料库中所属的文本类型、所在的文本、所在的行、行中的位置。配码就是前面提到的通过查询预先编好的词类词典,把可能的词类码配给每一个单词的过程,在 CLAWS 中是由一个称为 WORDTAG 的程序来完成的。需要指出的是,WORDTAG 在给词或标点配码时是每个词单独考虑的,而不考虑其上下文。还有,为了缩小词典的编辑量,这个系统还编了一个后缀表(suffix list),把一些能标志词类的后缀编入表中,如:-ness 标志着以它结尾的词一般是名词,-able 标志着以它结尾的词一般是形容词等等。没有编入词典的词就可以到这个后缀表中查找。此外还有许多其他细节问题需要考虑,限于篇幅不再详述。第三步,习语赋码是非常关键的一步。这一步是由一个称为 IDIOMTAG 的程序来完成的。在 CLAWS 系统中,它被看做是 WORDTAG 的一个有益补充,用于对一组一组的词进行赋码,以去除一些明显的错误。这个程序能在文本中寻找特定的词组,并作出相应的修改。例如:一旦它找到 as X as 这个

词组，而且 X 的词类中有 JJ（形容词），那么该程序就在第一个 as 后附上 QL（限定词），第二个 as 后附上 IN CS@（介词或很少情况下从属连词）。如果没有这个程序，那么 WORDTAG 会给每个 as 都附上三个词类码，这会给下一步的处理带来很大麻烦。所以 IDIOMTAG 的作用不可低估。接下来的是解决歧义码，上文中已经较为详细地对它进行了介绍，它是由称作 CHAINPROBS 的程序来完成的，主要任务就是通过上下文确定各词类码的可能性，一般情况下，可能性最大的词类就是正确的词类码。最后一步后编辑就是人工检查 CHAINPROBS 计算的可能性最大的词类是否是正确的词类，并去除多余的词类码。

关于词类赋码当然还有其他方法，限于篇幅，在此不作介绍（参见陈建生 1998；冯志伟 1997）。此外，我们在附录中列出了许多目前使用中的词类赋码系统。在下面一个小节中，我们再来看看 UCREL小组的句法赋码方法。

1.2.2.2　句法分析（parsing）

句法分析又称为句法赋码，就是对文本中的每一个句子进行句法标注。这里主要介绍 UCREL 小组的概率句法赋码系统。他们的句法赋码系统和词类赋码系统原理基本相同，而且句法赋码建立在词类赋码基础之上，即：词类赋码的输出正是句法赋码的输入。UCREL 的句法赋码系统主要分三个步骤。第一步是对文本中每一个词赋以可能的句法符。该步骤主要依赖于一部标明每一可能词类码对子（pair of tags）的句法符的词典。第二步是寻找一些特殊的语法码形式（patterns of tags）和句法片段（parse fragment），并对句法结构作必要的修改。最后一步是完成每一可能的句法分析，并逐一赋值，从中选出可能性最大，即值最大的句法分析作为每句的分析结果。现以该系统对句子 *I hope it's the right thing to do*. 为例说明句法赋码的过程（参阅 Garside & Leech 1987）。

CLAWS 赋码系统对该句赋以词类码后的结果如下：

P13	131	001			...
P13	131	010	I		PP1A
P13	131	020	hope		VB
P13	131	030	it	>	PP3
P13	131	031	's	<	BEZ
P13	131	040	the		ATI
P13	131	050	right		JJ
P13	131	060	thing		NN
P13	131	070	to		TO
P13	131	080	do		DO
P13	131	081	.		.
P13	131	082			...

第一步，对每个词赋上可能的句法符。为此，首先要编句法符词典。该词典不是以单个的词类码为条目的，而是以两个连续的词类码对子为条目的。UCREL 小组认为一个连续的词类码对子可以部分地确定某一成分的开始、继续或结束。例如，假设词类码对子是 noun-verb，它往往预示着一个名词词组的结束和一个动词词组的开始，用符号表示就是 N］［V；同样地，如果对子是 adjective-noun，它往往预示着一个名词词组的继续，用符号表示就是 N N。CLAWS 系统一共使用了 133 个词类码，如果要将每一对具体的词类码对子，即(tag^1, tag^2)编入词典，那么就意味着有 133^2 个条目。研究者们认为没有必要把这 133^2 对，即 17 689 个词类码对子全部编入词典，因为很多对子所预示的可能句法符是相同的。因此，在实际的编辑中就需要进行一些概括，以所谓的包含符(cover symbol)对子为词条，即(cover^1、cover^2)，只有有例外时才把具体的词类码对子(tag^1、tag^2)编入词典。例如，名词在 CLAWS 中被细分成好几十个类别，但用了包含符后就只有三种，即：N *，* S，* $，

分别代表单数名词,复数名词和名词的所有格形式。这就大大减少了所要编辑的条目,同时使系统更加有效。研究者们还发现,(cover¹、cover²)这样的条目并不是最简单的,因为有时候当(cover²)确定时,不管(cover¹)是什么,可能的句法符是不变的。在这种情况下,条目可以更加简化,即只考虑第二个包含符(cover²),例外的就编入(cover¹、cover²)。这样一来,词典实际上就由三部分组成,即(cover²)、(cover¹、cover²)、(tag¹、tag²)。这也就是说,第一部分给出的是不须考虑第一个包含符就能确定句法符的第二个包含符;第二部分是第一部分的例外,给出的是必须同时考虑前后两个包含符才能确定句法符的包含符对子;第三部分是第二部分的例外,即必须同时考虑前后两个具体词类码才能确定句法符的词类码对子。查词典是反向进行的,即:首先应该去查第三部分,如果查到了就可直接赋上相应的句法符;如果没有查到,则需要把(tag¹, tag²)转化成(cover¹, cover²),并到第二部分中去查,赋上相应的句法符;如果在第二部分还没查到,最后才到第一部分中去查,赋上相应的句法符。

就上述例句而言,需要在词典上查找的是(一、PP1A)、(PP1A、VB)、(VB、PP3)等词类码对子。有些对子只有一个可能句法符,有些有一个以上的可能句法符。句法符是由开始和结束方括号(即:左右向方括号)和句法成分标签构成的序列,这种序列也被称作 T-tag,所以通过查词典赋句法符也被称作 T-tag 赋码。一个 T-tag 由左边和右边两部分组成:左边部分表示当前的成分是什么,是否什么成分都可能,是否可能后跟一个或多个结束方括号;右边部分一般表示一个或多个新成分的开始,表示某些成分将继续,或者在极少的情况下表示一个新的,但尚未清楚是什么成分的开始。这个不清楚的成分在后来的分析过程中可以推断出来。以(VB, PP3)(动词的基形和代词 it)为例,它有两个可能的句法符,即:"Y][N"或"Y][Fn[N",这里,右向方括号表示成分的开始,左向方括号表示成分的结束,大写字母表示成分的类型,小写字母表示成分的细分类型(更详细的解释请参阅 Sampson 1987)。在这两个可能中,前者表示结束当前的成分,并开始一个名词词组;后者表

示结束当前的成分,并开始一个名词性从句,名词词组作该从句的第一个成分(这里的 Y 是可以作任何成分的通配符)。查词典的结果先赋在词类码对子的第二个词类上,有超过一个的句法符时用冒号分开(如下表的第二列所示)。然后对这些句法符进行调整,即把 T-tag 的左边部分赋在对子的第一个词类码上,把右边部分赋在对子的第二个词类码上并舍去通配符和重复成分,对有超过一种的句法符也是用冒号分开并加上大括号(如下表的第三列所示)。

…		
PP1A	Y [S[Na	[S[Na]
VB	Y] [V	[V]
PP3	Y] [N : Y] [Fn[N	{: [Fn} [N]
BEZ	Y] [V	[V]
AT1	Y] [N	[N
JJ	N N	
NN	N N : J] [N	{:]}
TO	Y [Ti[Vi : Y] [Ti[Vi	[Ti[Vi
DO	V V	Vi]
.	Y] S	S]
…	S]	

上表中第三列多数句法符是容易理解的,但 NN 所在的行中所赋的{:]}需要稍作解释。这行中第二列所赋的是对子(JJ、NN)的句法符有两种可能,即名词词组的继续或者是形容词词组的结束和名词词组的开始。前面的句法符中没有开始一个形容词词组,因此第二种可能,即要结束一个形容词词组与前面的句法符有冲突,可以排除,也就是说只有 NN 这个句法符是可能的。下一行(TO 所在的行)有两种可能,即延续或结束当前的成分,属于 T-tag 的第一部分被调整到前一个词类码上,{:]}就代表或者什么都不赋或者赋上结束当前成分的左向方括号。这也就是说,有一些句法符是可以直接排除的,但是仍

然有一些句法符需要作进一步选择。从第三列看，还有两处需要进行选择。这就是说到目前为止，我们只能认为该句有 4 种可能的句法分析结果，还需要作出选择。至此，句法赋码已经完成了第一步。为便于理解，在讨论第二步之前，我们先讨论一下第三步。

可以看出，第一步的结果中已经包含了一些成分的类型及其开始位置，但是还有许多成分的结束位置尚未确定。因此，第三步的主要任务就是确定这些成分的结束位置。所有可能的位置都要加以考虑，至于哪个位置最合适就需要由概率的大小来决定。概率的计算主要考虑每个成分由更小的成分组成结构的可能性大小。仍以上述句子为例。为方便描述，把上句的词类码换成它们相应的包含符，并从第一步所得的结果中选择一种，比如说在两个选择点中都用第二个选择，那么可以得到以下结果（* 号为某一词类码的包含符）：

I	P*A	[S[Na]
hope	VB*	[V]
it	P*	[Fn[N]
's	BE*	[V]
the	DT*	[N
right	J*	
thing	N*	N]
to	TO	[Ti[Vi
do	DO*	Vi]
.	.*	S]

可以看出，第三列的句法分析是不完全的，因为 Fn 和 Ti 还没有确定结束位置。就 Ti 而言，它结束的位置只有在 Vi 结束符之后才是唯一合乎逻辑的，因此它的结束位置是容易确定的。但是，Fn 的结束位置从理论上讲可以是名词词组之后（或者 it 之后，或者 thing 之后），也可以是动词词组之后（'s 之后），还可以是现已确定结束位

置的 Ti 之后(do 之后)。这意味着 Fn 的结束位置又有 4 种可能,每一种可能当然都要考虑。以第二种为例,即 Fn 在 thing 之后结束,其句法分析的结果如下:

I	P * A	[S[Na]
hope	VB *	[V]
it	P *	[Fn[N]
's	BE *	[V]
the	DT *	[N
right	J *	
thing	N *	N] Fn]
to	TO	[Ti[Vi
do	DO *	Vi] Ti]
.	. *	S]

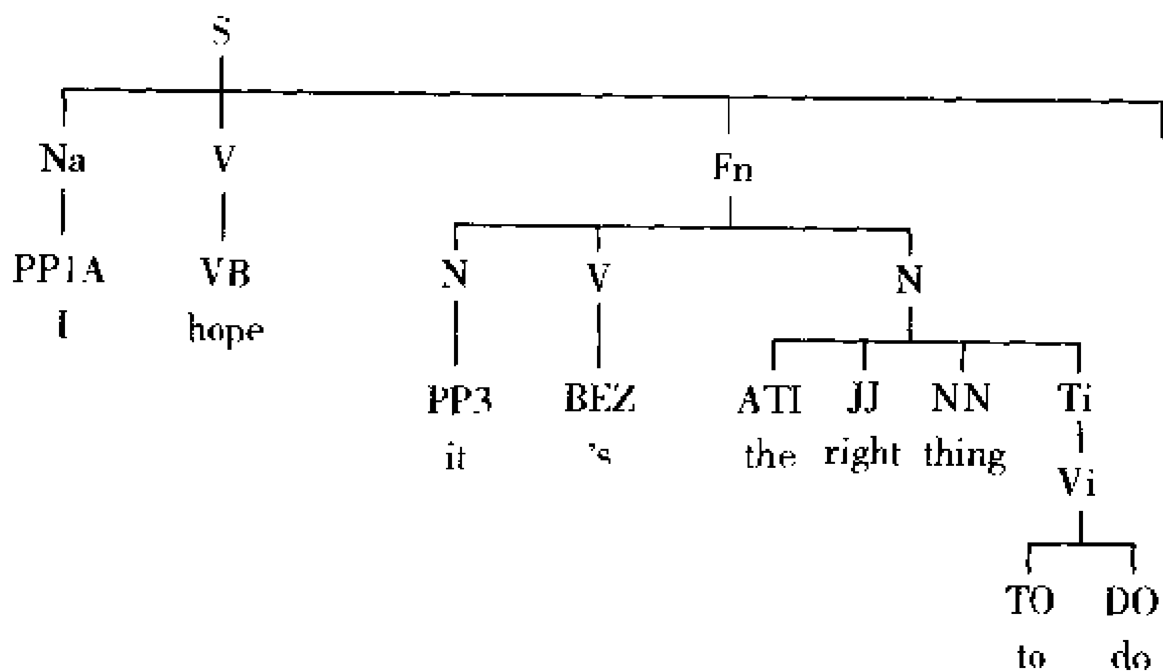
那么,这种句法分析的概率有多大呢? 概率是由一个所谓的"优点值"(figure of merit)来表示的,而这个优点值则以该句法分析中每一个成分以当前结构出现的概率为基础而计算出来。上述句法分析的优点值取决于以下各成分与其后的结构出现的概率:

S:	[Na V Fn Ti . *]
Na:	[P * A]
V:	[VB *]
Fn:	[N V N]
N:	[P *]
V:	[BE *]
N:	[DT * J * N *]
Ti:	[Vi]
Vi:	[TO DO *]

显然,这些概率信息需要从已赋句法码的语料库中提取。UCREL正是从语料库中已赋句法码的那一部分提取这些概率信息的。至于如何利用这些成分各自结构出现的概率计算优点值,可以尝试各种方法,例如将所有这些概率相乘,或者求它们的平均值等。计算方法要根据经验作出必要的修改。最后,程序选择优点值最大的句法分析作为该句的句法赋码,结果如下:

P13	131	001			...	
P13	131	010	I		PP1A	[S[Na]
P13	131	020	hope		VB	[V]
P13	131	030	it	>	PP3	[Fn[N]
P13	131	031	's	<	BEZ	[V]
P13	131	040	the		ATI	[N
P13	131	050	right		JJ	
P13	131	060	thing		NN	
P13	131	070	to		TO	[Ti[Vi
P13	131	080	do		DO	Vi]Ti]N]Fn]
P13	131	081	.		.	S]
P13	131	082			...	

最后一列就是该句的句法赋码,用树状结构可以表示如下:



至此，我们已了解了句法赋码系统的基本原理和方法。下面简要介绍上述句法赋码的第二步。

第二步也相当关键，它像词类赋码中赋习语码(IDIOMTAG)那一步一样，对第一步的结果作出必要的修改或补充，减少可能的句法分析，使第三步的负担大大减轻。正如 IDIOMTAG 对一些特定的词组进行特定处理一样，句法赋码中间这一步也是对一些特殊的语言结构进行特殊处理。例如，在遇到"介词 + 关系代词 which"的结构时，第一步的处理是到词典中去查(介词，wh-词)，而这个对于意味着某一类型的名词词组的开始。然而，在实际赋码中这是一个错误的决定，因为"介词 + 关系代词 which"意味着一个关系从句的开始。中间的一步便是在遇到该类问题时对结构赋上正确句法符。类似需要特殊处理的语言结构有很多，UCREL 在经验的基础上不断增加这些需要特殊处理的结构，这使第三步的处理负担不断减轻，从而大大提高系统的效率。

1.2.2.3 对当前赋码技术的评价

当前，自动词类赋码技术已经基本成熟，许多系统的准确率已达到和超过了 96%—97%，这样的精确度已经基本达到了实际研究和应用的需要。这样的系统都以语料库为基础，通过从语料库中挖掘出来的概率信息，对真实的语料进行赋码。这不仅充分证明了语言或语法的概率特性，而且为自然语言的处理开辟了广阔的前景。词类赋码和句法赋码为语言的量化研究创造了条件，为进一步研究自然语言的概率性特征提供了方便。词类赋码和句法赋码不是语料加工的终结，而是为进一步的语义和语用分析等打下了必要的基础。

词类赋码和句法赋码也存在一些问题。目前自动词类赋码和句法赋码系统有不少，但是，各系统的词类码和句法码不统一，使相关的研究不能共享数据，造成不必要的资源浪费。另外一个关键问题是词类码和句法码对词类及句法单位的划分没有客观的标准，划分可能会掩盖一些语言的本质特征。因此，对文本是否需要进行词类和句法赋码，语料库语言学家们的意见也不一致。此外，目前许

多句法赋码系统以词类赋码系统的输出为输入，把词类分析作为句法分析的低一层次的分析，这在一定程度上隔离了词汇和句法的关系。前面介绍的 UCREL 的词类赋码和句法赋码系统都有一个中间环节 IDIOMTAG，用于处理一些特定的词组，而句法赋码的中间环节是对一些特定的语言结构进行特殊处理，两者在表面上看只是增强系统有效性的一种手段，但实际上有意无意地把词汇和句法结合了起来。实际上，这两个中间环节在上述的系统中还远远没有发挥它们应有的作用，也就是说词汇和语法的密切联系还远远没有受到足够的重视。将来的赋码系统，尤其是句法赋码系统需要在它的中间环节上作更多的文章。

2 语料库的基本统计手段和原理

基于语料库的语言研究一般采取定性与定量相结合的研究方法。要进行定量研究就要涉及一些统计方法。这一节将就目前常用的统计方法及其原理作一简要介绍，以期使尚未进行过语料库研究的学者和语言教师对研究方法有一个认识。

2.1 文本总体统计特征

一个语料库往往由许多文本组成。前面已经提到，每一篇文本都是有标识的，是可以独立于语料库中其他文本的。因此，可以把每一文本作为一个独立的文件进行存储。当然也可以对语料库中的文本进行分类，把属于同一类型的文本放在一起作为一个文件进行存储。此外，也可以把整个语料库作为一个文件进行存储。

语料库分析软件一般对文件进行操作。如果每一文本是一个独立的文件，那么可以统计出每一文本的总体统计特征及整个语料库的总体统计特征。如果文本是分类存储的，那么可以统计出每一

类的总体统计特征及整个语料库的总体统计特征。如果语料库是一个文件，那么只能统计出整个语料库的总体统计特征。

主要的统计特征有：文件的字节数(bytes)、形符数(tokens)、类符数(types)、类符形符比(type/token ratio)、标准化类符形符比(standardised type/token ratio)、平均词长(average word length)、句子数(sentences)、平均句长(sentence length)、句长标准差(standard deviation of sentence length)、段落数(paragraphs)、平均段落长(paragraph length)、段落长标准差(standard deviation of paragraph length)等等。这些统计特征中的大多数很好理解，需要在此稍作解释的是形符、类符以及它们之间的比率。所谓形符数是指文本一共有多少个词，而类符数是指文本一共有多少个不同的词形。单纯的形符数和类符数不能反应文本的本质特征，但两者的比率却在一定程度上反映了文本的某种本质特征，即用词的变化性。例如，根据 Wordsmith 的计算，BROWN 语料库的形符数是 1 015 530，而类符数是 42 709，类符形符比为 4.21%。但是，该软件对 CLEC(中国学习者英语语料库)中大学英语子语料库的统计表明，该子语料库的形符数是 320 733，而类符数是 8216，类符形符比为 2.56%。从这两个语料库的类符形符比来看，本族语者的用词变化性明显比中国学习者要高。但是，英语的词汇量是一定的，如果语料库的容量不断扩大，形符数将随之扩大，然而类符数的增加却不可能保持同步，所以当语料库的容量达到一定程度时，类符数的增加将越来越小，类符形符比也将越来越小，两者的比率就无法反映用词的变化性。比如有一篇文本，它的形符数是 1000 个，类符数是 400 个，类符形符比是 40%。但是，这两者之间的比率随着文本长度的变化而有很大的变化，对于 1000 词的文章，这个比率可能是 40%；对于更短的文章，这个比率可能会上升到 70%；对于 400 万词的语料库，它的整体类符形符比可能只有 2%，如此看来，这样来计算两者的比率意义是不大的。因此，语料库语言学家们就采用了标准化类符形符比来反映用词的变化性。标准化类符形符比的计算方法是按一定的长度分批计算文本的类

符形符比，然后求出它们的平均值。如果将这个长度界定为1000，那么就先计算第一个 1000 词的类符形符比，然后计算第二个 1000 词的类符形符比，如此计算下去一直到文本的结束(最后所剩未满 1000 词的语料就不予考虑)，最后计算它们的平均值就是标准化类符形符比。这样计算出来的比率就可以用来比较不同长度文本的用词变化性，弥补了未经标准化的类符形符比的缺陷。类符形符比还可以用对数表示(参阅桂诗春、宁春岩 1997:141－143)。

2.2 词频统计

词频统计是语料库研究的一个基本的统计手段。早期的语料库研究基本上仅限于词频的统计。在机读语料库(machine-readable corpus)问世之前，研究者们是通过人工方法来计算词频的。用人工方法计算词频自然费时费力，且不够精确，但在当时条件下这些研究者们不惜花费大量时间和精力从事机械繁琐的词频统计，足见词频统计的必要性和重要性。词频统计的结果可以用于其他更为复杂的统计之中，如下面将要介绍的搭配词的计算中需要用到词频信息。关键词及关键性的计算更需要以词频统计结果为依据。

词频统计需要的实际上是一个简单的程序，多数语料库分析软件都有词频统计的功能。做词频统计后，一般可以产生两个词频表，一个以词的字母顺序排列，一个以词的频率大小排列。这两个词频表各有所长，都是进一步研究所需要的。

需要在此特别指出的是，"词"本身是一个非常模糊的概念，不太容易定义。一般软件都是统计各词形的频率，因此词频统计与其说是在计算词的频率还不如说是在计算词形(word form)的频率。此外，还需考虑撇号(apostrophe)和连字符(hyphen)等，例如有连字符的词应当看成是一个词还是两个词需要事先确定。目前常用的软件 Wordsmith 允许用户在这些方面做出选择，自行确定词的定义。

此外还值得一提的是,Wordsmith 还提供了一个统计词丛(cluster)频数的功能。所谓词丛是指由两个或两个以上的词形构筑的连续的序列,按所含词形的多少分 2 词词丛,3 词词丛,4 词词丛等。举个简单的例子,英语句子 *It is one of the most entertaining and irresponsible novels of the season*. 取自 BROWN 语料库。句中含有的 2 词词丛有:*it is*、*is one*、*one of*、*of the*、*the most*、*most entertaining*、*entertaining and*、*and irresponsible*、*irresponsible novels*、*novels of*、*of the*、*the season*、其中 of the 的频数是 2,其他词丛的频数均是 1。如果对整个语料库进行词丛统计,那么有些词丛的频数会相当高。这些高频的词丛对研究比词更大的单位,如搭配,是很有价值的。有些研究者直接通过词丛来比较不同的文本或语料库。

2.3 搭配词及搭配力的计算

搭配(collocation)是当前语料库语言学研究的一个热点。许多语料库语言学家对搭配进行了广泛深入的研究,但是至今对"搭配"这个概念尚未有明确的定义。卫乃兴(1999)对搭配作了全面深入的探讨,并对它作了较为严格的定义,指出了搭配的一些主要特征(参阅本书第三章)。这里介绍语料库研究中如何确定搭配词(collocate)和计算搭配力(collocability)。根据辛克莱的定义,搭配是"两个或两个以上的词在文本中很短的距离内的共现"(1991:170)。这个定义相当宽泛,只考虑词的共现,不考虑词与词之间的语法关系。如果一个词在文本中与被研究词的距离在规定的范围之内,那么该词就是被研究词的搭配词。但是,文本中在规定范围内与被研究的词共现的词一般有很多,有的词只出现一至二次,有的却出现多次。出现次数少的词之所以和被研究词共现可能并不是因为它们与被研究的词有密切的关系,而可能是由偶然或随机因素所造成的;出现次数多的词之所以和被研究词共现可能不是由于偶然或随机因素造成的,而可能是由它们与被研究词的密切关系所

决定的，即它受到了被研究词的吸引。如此看来，搭配还有一个是否有意义或有显著性(significance)的问题：如果两个词纯粹是由于偶然或随机因素而共现，那么它们之间的搭配没有多少意义；相反，如果两个词的共现不是由偶然或随机因素引起，而是两者存在密切关系，相互吸引的结果，那么它们之间的搭配是有意义的。出于这种考虑，有些语料库分析软件，如 Wordsmith、Mconcord，就把共现不足于一定次数的词不作为有意义的搭配词考虑．Wordsmith 的缺省值是 5 次，而 Mconcord 的缺省值是 3 次。

但是单从共现的次数看两个词的搭配是否有意义还不能准确反映实际情况。比如一个词可能和被研究词共现多次，但这并不能说该词一定是一个有意义的搭配词。这是因为，该词与被研究词共现可能并不是受被研究词的吸引，而只是由于该词在整个文本中非常常用，它与被研究词的共现只是出于随机的因素。那么，就不能单从共现的次数来看搭配的意义而应同时考虑该词在文本中的使用频率，由此也就产生了"搭配力"这个概念。搭配力大小反映的实际上就是搭配的意义或显著性的大小。搭配力越大，搭配的意义就越大；反之，搭配的意义就越小。搭配力一般以 Z 值(Z-score)或 T 值(T-score)表示(这两者在样本大的情况下，值相差很小)。下面将举例说明一下 Z 值的计算。假设被研究的词是 *conclusion*，所使用的语料库是 BROWN 语料库。被研究词 *conclusion* 在语料库中共出现了 56 次，如果它的出现次数以 N 表示，那么，N = 56。为了更好地说明问题，下面列出其中的 16 行索引，每行取节点词 conclusion 前后各 4 个词为其搭配词：

1	do not warrant this	conclusion. There are increasing number
2	tube will be listed. The	conclusion shall be reached that
3	Gradually he reached a	conclusion. The marine was alone
4	which led to this	conclusion, the Court remanded the
5	This leads to the	conclusion that "the fact is
6	1959 ~ 1960 leads to a	conclusion that desegregation-from-court

7	workers lead to the	conclusion that the * * f bond
8	Pike jumped to the	conclusion that Robinson was guilty,
9	one can escape the	conclusion that the Negro's status
10	hard to escape the	conclusion that the softening process
11	to a dangerously erroneous	conclusion about the West's ability
12	come to the same	conclusion. What they meant was
13	at substantially the same	conclusion, but skirts the problem
14	has arrived at this	conclusion, for it means that
15	with mine. This second	conclusion, independently arrived at by
16	first arrived at this	conclusion. She thought it was

这前后的 4 个词就是定义“搭配”时提到的跨距(span)范围,以 S 表示,即 S = 4。如果把所有的索引行都列出来,就像是在 *conclusion* 的周围形成了一个小的文本,语料库语言学家把这个小的文本称作小文本(mini-text),以 M 表示,则 M = (2S+1)×N = (2×4+1)×56 = 504。按照 Wordsmith 对 BROWN 语料库的统计,该语料库的大小是 1 015 530 词。把语料库当作一个文本,它便是整个文本的长度,以 W 表示,则 W = 1 015 530。从以上索引行可以看到出现在 *conclusion* 周围的词有不少,其中有些词出现一次,有的则出现两次,有的两次以上。本文中用括号中的数字表示它们在跨距内出现的频数或次数,并以 *warrant* (1)、*reached* (2)、*leads* (2)、*escape* (2)、*arrived* (3) 这些与 *conclusion* 搭配的动词为例,说明它们与 *conclusion* 的搭配力。

前面已经提到,在计算搭配力时需要考虑搭配词本身在整个文本的出现频数(或数次)。例如,动词 *leads* 在整个文本中共出现 33 次。如果以 C 表示出现频数,C = 33。知道了搭配词 *leads* 在整个文本的出现次数 C 和整个文本的长度 W,就知道了它在整个文本中所占的比率。如果以 P 表示比率,则 P = C / W = 33/1 015 530 = 0.000 032 5。这也就是说,平均在每 100 万词中,该词出现 32.5 次。

如果 *leads* 的分布是均匀的或者是随机的,即不受 *conclu-*

sion 的影响，那么在 *conclusion* 周围以索引行形成的小文本中，它被期望出现的次数接近于 P×M，这个数值称为期望频数。

如果以 E 表示期望频数，则 E = P × M = 0.000 032 5 × 504 = 0.0163 8。这也就是说，人们预期它在小文本中出现 0.016 38 次，即远远不到 1 次。然而，事实是 *leads* 在小文本中出现了 2 次。如果我们以 C′表示出现的实际次数，C′ = 2。实际值与期望值之间的差距，以 Z 值来体现。在计算 Z 值之前，先要计算标准差 SD，它的计算公式是：

$$SD = \sqrt{P \times (1 - P) \times M}$$

按照这个公式，*leads* 在小文本中分布的标准差是：

$$SD = \sqrt{P \times (1 - P) \times M} = \sqrt{0.000\,032\,5 \times 0.999\,967\,5 \times 504} \approx 0.128$$

知道了标准差，就可以计算 Z 值了，它的计算公式是：

$$Z = \frac{C' - E}{SD}$$

按照这个公式，*leads* 作为 *conclusion* 的搭配词，它与 *conclusion* 的搭配力是：

$$Z = \frac{C' - E}{SD} = \frac{2 - 0.016\,38}{0.128} \approx 15.500$$

如果 Z 值达到 1，就可认为有一定的意义，但是为了更保险起见，研究者通常把 Z 值定为 2，即 Z 值超过 2 的搭配词被认为是有显著意义的搭配词。上述其他动词 *warrant*、*reached*、*escape*、*arrived* 在整个语料库中分别出现 20、170、64 和 62 次，计算后所得 Z 值分别是 9.938、6.596、11.044、16.927。这些值都比 2 大得多，因此都有显著意义。可以看到，虽然 *warrant* 在小文本中只出现了 1 次，但它与 *conclusion* 的搭配力已经达到 9.938，它与 *conclusion* 的共现已经不能以偶然或随机因素来解释了，我们只能把这种共现归

因于 *conclusion* 对之的吸引。另外，虽然 *reached*、*leads*、*escape* 在小文本中出现次数相同，但由于它们在整个文本中出现的频数不同，它们的 Z 值也各不相同，在这三个词中 *conclusion* 对 *leads* 的吸引最大，*escape* 次之，*reached* 最小。可见搭配力用 Z 值来反映更加科学。Z 值的计算过程可以总结如下：

要计算 Z 值，需要知道 5 个数据，它们分别是：被研究词或节点词的频数 N，跨距 S，搭配词在整个文本中的频数 C，搭配词在小文本中的频数 C′以及整个文本的长度 W。根据这些数据，可以计算小文本的长度 M，搭配词占整个文本的比率 P，搭配词在小文本中的期望频数 E，其标准差及最终的 Z 值或 Z 分数。它们的计算公式分别是：

$$M = (2S + 1)N$$
$$P = C / W$$
$$E = P \times M$$
$$SD = \sqrt{P \times (1 - P) \times M}$$
$$Z = (C' - E) / SD$$

Z 值是一个很有用的统计量，有些语料库分析软件，如 TACT 就提供了搭配词与节点词的 Z 值，用来揭示搭配力。

除了上述的统计手段之外，另外一个非常有用的统计手段是关键词及关键性（key-ness）的计算，下一小节将讨论这个内容。

2.4 关键词及关键性的计算

这里所说的"关键词"与 KWIC（key word in context）中的 key word 不是一回事。后者是作索引时的节点词（node word）或搜索词（search word）。而前者则是指文本中的关键词，指的是跟某一标准相比其频率显著偏高的词，偏高的程度就是该关键词的"关键性"。

按照上述定义，一个词是否是某一文本或文类(genre)的关键词，不仅取决于该词在该文本或文类中的出现情况，还取决于该词在与之相对比的参照语料库中的出现情况。举个例子来说，假如定冠词 *the* 在某一长度为 1000 词的文本中出现了 50 次，其出现频率达到了 5%，但不能因此说 *the* 是该文本的关键词，这是因为 *the* 在任何文本中的频率都很高，不是唯独在这一文本中出现频率高，它在参照语料库中的频率可能还不止 5%，单就其在该文本或文类中出现的频率来决定它是否是关键词显然是不合适的。

参照语料库在长度上要求比与之相比的文本或文类长，它往往反映了词在一般情况下在语料中出现的情况，因而可以作为比较的标准。比如，假设 detective 一词在参照语料库中出现的频率为 1/10 000，而该词在与之相比的文本中的出现频率却达到或超过了 1/100，我们就有理由相信该文本跟 *detective* 很有关系，很可能是个侦探故事，这是因为 *detective* 的频率与一般出现频率相比在该文本中明显偏高。关键词频率的偏高程度可用两种方法计算，其一是 X^2 值，其二是对数或然率(log likelihood)。这里只介绍 X^2 的计算。X^2 的计算方法并不复杂，它牵涉到四个变量，即某词在文本中的频数 a，所在文本的长度 b，该词在参照语料库中的频数 c，参照语料库的长度 d。对这四个数作独立样本表格，带亚茨连续性校正的 X^2 检验所得的 X^2 值就是关键性的值。计算公式如下：

$$x^2 = \frac{\left(|ad - bc| - \frac{N}{2}\right)^2 \times N}{(a+b)(a+c)(b+d)(c+d)}$$

(注：N 是 a、b、c、d 的总和)

为了更好地说明问题，仍以上述 *detective* 一词为例说明关键性的计算过程。假设它在被研究文本中的频数是 10 次，而被研究文本的长度是 1000 词，在参照语料库中的频数是 1 次，而参照语料库的长度是 10 000 词，那么就可以作出如下的表格：

	词的频数	文本长度	总　和
被研究文本	a = 10	b = 1000	a + b = 1010
参照语料库	c = 1	d = 10 000	c + d = 10 001
总和	a + c = 11	b + d = 11 000	N = 11 011

把以上数值代入上述公式,则:

$$x^2 = \frac{\left(|10 \times 10\,000 - 1000 \times 1| - \frac{11\,011}{2}\right)^2 \times 11\,011}{(10 + 1000)(10 + 1)(1000 + 10\,000)(1 + 10\,000)}$$
$$= 78.7$$

像上述的情况,X^2 值在一般的统计中只要大于 3.84,已经有显著意义。现在 X^2 的值已经达到 78.7,远远高出 3.84,其显著意义也就显而易见了。这个 X^2 值就是 *detective* 在被研究文本中的关键性。关键性还应参照相应的 P 值,以确定其显著性水平。P 值越低,显著性水平越高,作出判断的把握越大。一般情况下 P 值小于 0.05 就有显著意义,但在关键性的计算中 P 值一般可以定得非常小(如 Wordsmith 默认的最大 P 值是 0.000 001),这样可以避免不够关键的词被算作关键词。

关键词的自动提取对语言研究与应用的价值正在探索之中,其潜在的应用价值是相当可观的。

结　束　语

语料库语言学作为一门年轻的科学还有极大的发展潜力。大规模的语料库等待着研究者们去建设和研究。在建设语料库时,需要考虑和解决语料库的代表性问题。不断扩大语料库的规模是增强语料库代表性的有效方法之一,但语料库的内容、各类语料在库

中的比率也是应当考虑的重要问题。在语料库基础上开展的研究大大提高了人们对语言的认识水平,这也必将有助于更好地解决语料库的代表性问题。语料库的预加工为更深层次的分析创造了条件。经过预加工的语料为语言研究提供了方便。

通过对语料库的研究,研究者们已经逐步形成了一些基本的统计手段和方法。了解和掌握这些统计方法及其原理对从事语言研究的学者和教师不无裨益。

语料库研究需要的是一个开放的思路,而不是固步自封。应不断去尝试和探寻更好的研究方法和手段。

第五章 文本索引工具及应用

1 文本索引

90 年代,用于外语教学的各种语料库由于计算机技术和因特网的发展得到快速发展和传播。各个国家和地区纷纷建成或正在筹建各种通用和专用语料库,并广泛应用于外语教学和研究。我国语料库建设起步于 80 年代中期,以上海交通大学建成的科技英语语料库为标志。近两年来,中国英语学习者语料库建成并完成人工赋码,标志着我国应用于外语教学与研究的语料库发展进入了一个新阶段。但是,广大外语教师对这一有力工具尚感陌生,尤其对语料库索引软件在外语教学中的应用感到无从下手。一些偏重语料库理论的文章仅限于西方现有的相关成果的介绍,没有论述语料库在我国外语教学与研究中的实际应用,外语教师对语料库这一先进技术仍“望库兴叹”。实际上,语料库技术及其索引软件是一项应用性很强的外语教学辅助工具。本章从应用入手,通过常用的语料库索引软件 MicroConcord 应用分析示例,探讨外语课堂教学中实时运用语料库索引的可行性和基本手段,以使外语教师对使用语料库尽快上手,使之服务于自己的外语教学。通常意义上的计算机语料库一般分为语料库本体(即语料库电子文本)和语料库引擎(即语料库索引程序)。二者可分别获得。某些现有的索引软件对所处理的电子文本在格式上有特殊要求。外语教师也可以动手建立自己的专用语料库,并利用现有的索引软件进行查询。(有关语料库建设问题详见本书第二章和第四章的讨论。)

2 索引的意义

文本索引在文本分析中的应用具有悠久的历史传统。西方最初的词语索引可追溯到中世纪。早期在对宗教文件进行文本分析中,以字母顺序排列伴随有限的语境词的索引词表开始出现,如1387年特里维萨(Trevisa)对圣经文本的词语索引(Ball and Taylor 1995)。1845年,克拉克(C. Clark)夫人对莎士比亚的5部戏剧作品进行了手工索引,大约花了16年功夫,其主要工作是索引出不带语境的关键词,如下所示:

SEIZETH but his own	*Titus Andronicus*, i. 2
SEIZING him, the benefit	*Richard III*. iii. 1
SEIZURE, do we seize into	*As you Like it*, iii. 1
unyoke this seizure and this kind	*King John*, iii. 1
to whose oft seizure the cygnet's	*Troilus & Cress*. i. 1

(引自 Ball and Taylor 1995)

现代的索引软件开始于本世纪50年代末,开发这种程序的目的是利用科技论文标题及论文摘要中的关键词对文本进行分类,以便索引和查询。

文本索引作为一种强有力的文本分析工具,近年来被广泛应用于三个主要领域:语言学、文学以及语言教学。在语言学研究中,文本索引为句法学、语义学、语用学以及词语研究提供了可靠的量化依据。通过观察大量真实语言材料,可以分析不同的句法结构的分布,也可以对自然语言中存在的大量的一词多义(polysemy)和同形异义现象(homonymy)进行区分。由于索引行基于连续的篇章,通过扩展语境也可以分析某一话语结构中主题词之间的复杂关系。在研究词语搭配关系时,可观察的语境

往往是与某个词相邻近的（contiguous）或非邻近的搭配词（collocates），其分析范围属于短语或句子层面，所研究的是词语组合的线性关系；而在对单个语篇所进行的分析中，对词语的研究可以扩展到整个语篇，进而研究围绕某一主题词语的非线性联想关系。在文学作品研究中，可以通过文本索引对某一作家的作品集，或对单个作家某一文体的作品进行分析和对比研究。这种分析对研究作家某种语体风格的形成或主题的表达提供可靠的量化依据。文学批评中基于作品文本的分析研究避免了只注重概念的演绎，或生搬各种文学以外的理论进行附会穿凿的流弊。事实上，不少索引软件开发的最初目的就是用来进行文学作品分析的（如TACT、Concordance 等）。在语言教学中，文本索引在课堂教学中的应用为验证语法规则、词语用法以及词语搭配规则提供了方便的工具。如凯特曼（Kettemann）博士通过语料库索引对英语中的间接引语进行了观察。（见以下索引行。）夸克等（1985：1026）对间接引语的陈述为：1）直接引语的现在时在间接引语中转换为过去时；2）直接引语的过去时转换为过去完成时；3）直接引语的现在完成时转换为过去完成时。在索引行中，他发现那些表示事实、意图以及最近发生的间接引语中，时态的运用与规则不相一致。他认为，外语教学不应过分注重形式规则，而忽略了意义和语境。在教学中，至少要向学生解释清楚为什么会有这么多例外以及这些例外所表达的意义（Kettemann 1996）。

1 aspects" in his case, but has not said he deserves asylum. <p> About 3500 Kurds came
2 er the elections. South Africa has said it expects the UN to reduce its 4000-strong
3 Chile (AFP) President Pinochet said he has decided to abolish the secret police
4 nvestment previously. <p> Mr Young said he has had to sell stock he would prefer
5 diary of BAT Industries, yesterday said it has filed a motion in Los Angeles County
6 on a new lease. <p> Mrs Aquino has said she is "keeping her options open" on
7 a revenue stream for BT. Mr Booth said that BTI is concentrating more and more
8 be 30 000 tons. The Russians have said they have 50 000 tons. <p> But such

9 group that is up for sale and has said it wants to sell AMI. Bupa and Sun Alliance
10 said, "Who would believe us if we said we are changing our policy to win votes?"

以上索引行中,第1、2、9行在间接引语从句中都使用一般现在时,是对处于一般状态的心理活动的转述,如主观评价和意图。第6、7、10行在间接引语中使用了一般现在进行时,表明正在发生的动作。第3、4、5、8行则使用了现在完成时,表示到转述时间为止已经完成的动作或已经具有的状态。

对于外语教学中语法和词汇教学而言,应用语料库索引的优势是显而易见的。语料库索引为语言教学提供真实可靠的语境信息,使教师和学生在真实语言应用中验证教科书和词典中所给的定义和解释,从而使学习的过程变成自我探索和自我发现的过程。在外语教学中,语料库索引在语法和词汇教学与学习中的应用有以下几个方面:1)通过对语法结构、时态、功能词等检索,可以即时验证各种语法结构的典型用法。这种语法学习把要学习的语法特征在真实的语境中呈现出来,强化学习者的自我发现和语法意识培养,比传统的孤立的语法讲解和脱离语境的语法练习具有多方面的优越性。另外,语料库索引可以提供有关词汇用法和意义的真实信息,并以此检验词典或教科书中提供的解释和说明。通过索引,学习者可以体验词汇或词组在不同语境中的确切使用,以增加感性认识。2)同义词比较。语料库索引可对同义词群提供丰富的用法和语境,使学习者能够比较和掌握同义词之间细微的语义语用差异。3)词语搭配。语料库索引提供词汇搭配的频率信息和用法,平时靠教师语言直觉无法确定的问题可迎刃而解。4)编制即时课堂练习。多数索引软件提供搜索词遮蔽(Zapping)功能,使教师能够利用索引行轻松编制填空练习。

例如,学生对 *still* / *yet* / *already* 三个词的实际用法感到迷惑。尽管语法教材中提供了详尽的解释,但学生在应用中仍经常用错。教师可以通过词语索引,让学生在课堂上实时验证这些词的用法。通过使用 JDEST 语料库进行索引(限制索引数为 500),可以

看出,在这三个词中,*still* 最为常用(出现256次),其次是 *already*(143次),*yet* 出现次数最少(101次)。观察其应用语境,可以看出,*still* 右边最常用的词是"动词+ing+ for",最常用时态为现在完成时与现在完成进行时。观察搜索词左边第一个词,可以看出 *still* 和 *already* 经常出现在 *be* 和 *have* 的现在式和过去分词中间,而 *yet* 经常和 *not* 连用(参见 Murison-Bowie 1993)。

同样,利用词语索引也可以检查英语中同一个词不同的拼写形式。下面以 *realize* 和 *realise* 为例,检索 JDEST,BROWN,和 LOB 三个语料库,观察这两个拼写形式的出现频率,结果如下表:

	JDEST	BROWN	LOB
realize	114	67	26
realise	19	0	39

值得注意的是,在 JDEST(上海交大科技英语语料库)中,*realize* 出现的次数远远多于 *realise*。而在由60年代初纯粹美国英语构成的 BROWN 语料库中,*realize* 出现了67次,*realise* 一次也没出现。相比而言,在由60年代初纯粹英国英语构成的 LOB 语料库中,两种拼写形式几乎平分秋色。这表明 *realize* 是一个美国式拼写,并于60年代已在英国英语中得到使用,与 *realise* 这种形式并行存在。

3　索引工具的基本功能

索引工具的基本功能包括词表生成、语篇统计、"带语境的关键词"(KWIC)索引、排序、搭配词统计、词语型式统计、主题词提取、词丛统计、联想词统计及重组,以及词图统计。下面分别通过示例

介绍。应用工具包括 MicroConcord (Tim Johns & Mike Scott)、Wordsmith Tools (Mike Scott)、Concordance 1.1.3 (R. J. C. Watt)和 TACT。

3.1 词表与语篇统计

在对语料库文本进行统计分析中，词表功能和语篇统计功能把语料库中出现的所有“类符”统计列表。通常，词表提供以下几种信息：1）类符总数。词表中每个类符按字母顺序和频数顺序排列，可以进行顺序和逆序转换；2）每个类符的频数。类符的频数即该类符出现的总量。通过观察类符的分布，可以得到某一文类的语篇中词汇的基本分布信息，并通过编辑，提取常用词汇；3）每个类符的频率。由于不同语料库或文本的大小存在差异，不能简单地进行词汇频数比较。比如，*like* 这个类符在 BROWN 语料库中出现的频数是 1319，而在 JDEST 语料库中出现了 1852 次，不能据此就认为该类符在 JDEST 中比在 BROWN 语料库中更为常用。某个类符的频数由于语料库的大小不同，所代表的意义也不同。所以，在统计中还要计算类符的频数与语料库总量的比率，即该类符的频率：

$$\text{类符频率} = \frac{\text{类符观察频数}}{\text{语料库形符总量}}$$

上述的例词“like”在 BROWN 语料库中的频率为 0.11，而在 JDEST 中则只有 0.05，所以，该词在 BROWN 语料库中更常用；4)类符/形符比率。通过类符/形符比率统计，可以观察某篇文本或某个语料库的词汇密度(lexical density)；5)平均句长与段落长度；6)单词长度统计。词长即每个词的字母数量，工具软件统计出从单字母词到多字母词各自的频数，如下图是对英国著名作家 Oscar

Wilde 作品的平均词长所作的统计：

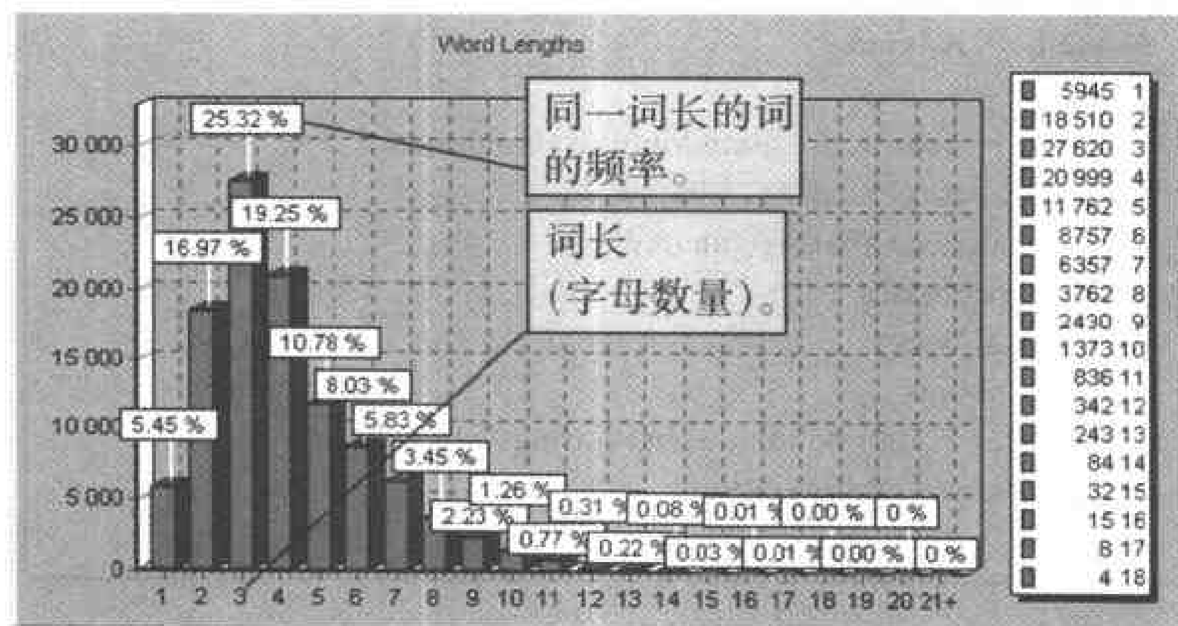


图 5.1　词长统计样例(Oscar Wilde 作品集)

词表与语篇统计为研究语料库或某个语篇的词语分布和词汇密度提供可靠依据。在有些软件中(如 Wordsmith Tools),还可以对词表中各个词形的屈折形式进行削尾处理(lemmatization),如把 *works*、*worked*、*working* 归为 *work*。在进行统计时,需要另外制作一个削尾处理文件,使用时调入该文件即可,如下图(图 5.2)是应用于 Wordsmith Tools 的削尾处理文件。该文件被调入后自动对词表进行削尾处理,并统计分布频数。

另外,词表还可以用来与其他语料库或文本进行对比,提取技术词汇或主题词。在外语教学中,语料库词表为确定教学大纲中的词汇范围以及选取尺度提供客观依据。

'nt → 't, nt, t
a → an
A-bomb → A-bombs
abacus → abacuses
abandon → abandons, abandoning, abandoned
abase → abases, abasing, abased
abate → abates, abating, abated
abbess → abbesses
abbey → abbeys
abbot → abbots
abbreviate → abbreviates, abbreviating, abbreviated
abbreviation → abbreviations
abdicate → abdicates, abdicating, abdicated

图 5.2 削尾处理词表(部分)(由 Yasumasa Someya 提供)

3.2 KWIC 索引

语料库最基本的分析手段是通过全文检索和词语索引(concordance)来实现的。词语索引的基本意义是把搜索词或词组按字母或频率顺序排列与其所在语境一同展示。词语索引最常见的形式称作 KWIC (key words in context),即"语境中的关键词"。"带语境的关键词搜索"(KWIC)技术为每一个搜索到的关键词提供所在行固定数量的语境词,并以该关键词为中心在屏幕上显示出来。索引软件在文学和语言学分析中的应用包括搭配分析、主题分析以及用于词典编纂的针对某一词汇的例句援引、语音分析、词素分析、词汇语义学和话语分析等。

关键词在不同的索引分析或软件中的其他名称有:搜索词(search word, 简称 SW),如 MicroConcord、Wordsmith Tools 等;节点词(node word) (如 Sinclair 1996)。为与本章以后讨论的"主题词"区分开来,在索引中用户键入的词统称为"搜索词"。以搜

索词为中心左右显示的词数构成了该搜索词的"跨距"(word span)。"跨距"中的词构成了搜索词的微型语境，或"共现语篇"(co-text)。该语篇是连续的文本，在索引中可以围绕任何搜索词、从搜索词所在行、段落以至整篇无限扩充显示。这种方法对单个的搜索词不仅提供短语和句子层面的使用语境，还可提供整个语篇。图 5.3 显示利用 MicroConcord 从上海交通大学科技英语语料库中搜索 *many a* 词组所展示的词语索引行。在课堂教学中，可以通过词语索引行，使学生观察词组 *many a* 与后跟的名词或名词词组数的一致性。该词组在形式上是单数，表达的却是复数意义。学生可以通过实例观察该词组应用语境。通过对词语索引行在屏编辑，可遮蔽搜索词，把索引行直接制作成填空练习，使学习者在真实的语境中掌握某一语法结构或词语用法。另外，基于语料库索引制作"数据驱动"学习课件，也是近年来课件开发中涌现的新事物(参见本书第二章)。该课件要求学生根据提供的语境猜出应填入的词。如学生难以猜出，再提供更多的索引行，这样，使词语学习过程成为一种发现和探索的过程。同时，也可以通过实时提示，以减少学习的难度。

1	ftsman can manage a wooden dial. and many a boy has found it an interesting and
2	boratory techniques, techniques that many a famous man came to admire, and for
3	n one which followed straight lines. Many a final chapter in works dealing with
4	r mother's companion and helpmate in many a household and nursery experience, fo
5	f the hole — instead of bolting, as many a lesser mule would have done when the
6	ral-parasitic trait, so typical for many a poet, was highly developed in him.
7	several thousand pounds a year, and many a professional man, say a doctor or ci
8	of land decision units. This is why many a public environmental measure has no
9	al and bureaucratic pressures. After many a sophisticated plan lay on the shelf
10	so as to bring out many a vein and many a tint, which escapes the eye of commo
11	en they meet for the first time. And many a tomcat has appeared with bloody nose
12	forms this truth so as to bring out many a vein and many a tint, which escapes

图 5.3 KWIC 索引示例(JDEST 语料库)

1	ftsman can manage a wooden dial, and	boy has found it an interesting and
2	boratory techniques, techniques that	famous man came to admire, and for
3	n one which followed straight lines	fina. chapter in works dealing with
4	r mother's companion and helpmate in	household and nursery experience, fo
5	f the hole — instead of bolting, as	lesser mule would have done when the
6	ral-parasitic trait, so typical for	poet, was highly developed in him.
7	severeal thousand pounds a year, and	professional man, say a doctor or ci
8	of land decision units. This is why	public environmental measure has no
9	al and bureaucratic pressures. After	sophisticated plan lay on the shelf
10	so as to bring out may a vein anc	tint, which escapes the eye of commo
11	en they meet for the first time. Anc	tomcat has appeared with bloody nose
12	forms this truth so as to bring out	vein and many a tint, which escapes

图 5.4 搜索词的遮蔽

在对词语索引行进行分析时,往往需要对要分析的语言现象分类统计。如分析某一个名词的用法时,可以对该名词大致分为以下几类:1)名词单独使用;2)名词词组;3)复合名词。在分析时我们希望把同一类别的用法放在一起集中观察,可以通过手工检查把每个词语索引行编号归类,再通过排序使相同类别的用法集中排列。编排好的词语索引可以存盘,待以后进一步分析。值得一提的是,大多数索引软件所提供的词语索引行是根据用户定义的跨距(如±5)以搜索词为中心显示结果,其语境并不完整。而以整句为索引单元的词语索引可以提供更完整的语境。但在这种词语索引中,搜索词不再显示在中心位置。

3.3 搭配词统计

在文本分析中,凡是处于跨距范围内的词都被看做是搜索词的共现词(co-occurring words),这些词与搜索词具有一定频率的共

现。根据共现词与搜索词的距离又分为邻近搭配和非邻近词项组合。通常意义上的搭配是指具有一定句法关系的搭配，如 *strong tea*，*bread and butter* 等，在这种搭配中，某一个成分的出现往往预示另外一个成分的存在(有关搭配问题的请参见本书第三章的讨论)。当然，仅凭共现词的频数并不能确定该词是否与搜索词具有真正意义上的搭配关系。一些高频共现词的出现可能是由于该词本身在语料库中频率就很大；另外一些共现词的出现很可能纯属偶然。在语料库统计中有两种方法常用来计算某一共现词与搜索词的搭配强度或搭配力(strength of collocability)。一种是通过计算 Z 值或 T 值表示搭配的强度，Z 值越高，搭配强度越大。另一种是通过计算在跨距内每个词位的频数分布，根据峰值的显著性来确定搭配强度(有关搭配的统计原理参见本书第四章的讨论)。如大学英语学习者语料库中 *knowledge* 一词的共现词与搜索词的搭配强度可以用以上两种方法统计出来。在表 5.1 中，每个共现词与搜索词的搭配的强度用 Z 值表示，并从高到低排序。在表 5.2 中，通过词位分布显示出各个共现词的搭配强度。

表 5.1 大学英语学习者语料库中 *knowledge* 的共现词

共现词	Z 值
learn	37.776
learned	19.395
books	18.256
enrich	16.070
enwide	15.225
enlarge	14.489
broaden	13.560
acquire	11.494
specialized	11.354

表 5.2 “*knowledge*”在大学英语学习者语料库中共现词的词位分布统计

N	Word	Total	Left	Right	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	Knowledge	588	13	19	1	1	4	7	0	556	0	7	6	3	3
2	The	298	177	121	32	20	12	20	93	0	0	57	19	22	23
3	And	184	6	118	13	9	28	8	8	0	66	14	9	13	16
4	Learn	179	144	35	5	20	14	87	18	0	0	5	7	14	9
5	Can	174	123	51	30	19	67	7	0	0	6	5	12	8	20
6	More	93	76	17	4	2	9	5	56	0	1	2	7	3	4
7	Our	79	63	16	6	6	2	5	44	0	0	4	7	0	5
8	From	76	6	70	3	2	0	1	0	0	33	4	10	11	12
9	They	66	37	29	12	20	3	2	0	0	2	6	6	13	2
10	But	64	23	41	4	6	6	6	1	0	7	6	10	12	6
11	Have	57	33	24	4	5	8	14	2	0	2	7	3	5	7
12	Society	56	28	28	8	8	8	2	2	0	0	9	12	0	7
13	That	56	16	40	7	4	1	3	1	0	19	3	5	7	6
14	You	50	22	28	3	14	4	1	0	0	4	8	8	3	5
15	Get	48	35	13	3	5	2	20	5	0	0	2	2	2	7

一般来说，通过 Z 值计算出的搭配强度信息更为准确。但是，通过词位分布统计可以更直观地观察各个词位上某一搭配词是否显著。如表 5.1 通过 Z 值排序，统计出学习者语料库中与 *knowledge* 搭配强度最高的词依次为：*learn*、*learned*、*books*、*enrich*、*enwide*、*enlarge*、*broaden*、*acquire*、*specialized*。而在词位分布统计中，只有搭配强度最高的词 *learn* 排在第四位；其他词都排在前十个以后。由此可见，词位分布统计虽然提供有关搭配词的位置信息，但同时也包含大量“噪声”信息(noisy information)，不能直接用作搭

配研究依据。

词语搭配分析对研究词语行为具有重要的意义。词语与词语搭配不仅对确立句法结构关系起着决定性作用，而且是意义呈现以及消除歧义的基本依据。其一，"词语像人类一样喜欢聚集"，一个词的出现往往预示或决定其他词的出现。其二，词意是通过共现词语的组合关系凸现出来的。脱离语境与搭配的孤立词语没有任何意义，词语只有在运用中以及与其他词语的搭配关系中才获得意义。其三，任何语言的基本构成成分是词语，而不是别的什么。所以，研究词语与词语搭配在句法学、语义学研究、以及语用学研究中具有重要价值。在外语学习中，学习者并不是孤立地学习单个的词汇，而是成组成块地学习和运用。

3.4 词语型式(pattern)统计

词语型式统计即根据索引中所规定跨距，计算在跨距内每一个词位上每个词语出现的频数，并按频数降序排序。词语型式统计直观展示某一搜索词前后各个词位的词语分布情况。根据词语型式统计，可以观察词语搭配关系以及围绕某一搜索词不同的聚集词群。这种方法与上述的词语搭配统计互为参照。下表显示的是大学英语学习者语料库中搜索词 *knowledge* 的词语型式统计。从表中可以看出，左一词位上(L1)频数从高到低依次排列的词大多是限定词，其中 *many* 一词可能是一不正确的搭配。根据该线索，再以 *knowledge* 为搜索词进行索引，以 *many* 为语境词(context word)，限定跨距为左一词位，进一步观察这一错误搭配的具体语境。左二词位上的词大多是与搜索词搭配的动词，如 *learn*、*get*、*use*、*know*、*have*、*improve*、*master*、*study*、*enrich* 等。靠近搜索词的词位上的词可能与搜索词有明确的语法搭配关系。比如左五词位"*think*"可能在其他统计中也会列入搭配词表中，但由于其位置太远，基本上可以排除它与搜索词的类联接关系，即动名搭配结构的 *think knowledge*。从表中可以看出，学习者使用的与

knowledge 搭配的动词属于随意性词语组合，其中很多搭配并不符合英语本族人的语言习惯(参见以后分析)。

表 5.3　搜索词 *knowledge* 的词语型式统计

N	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
1	The	they	can	learn	the		and	the	the	the
2	Can	learn	and	lot	more		from	books		learn
3	And	the	only	get	our			and	can	they
4	They	can	learn	the	new		that	for	books	and
5	Society	you	should	use	much		which	skill	from	but
6	That	campus	the	know	some	well		practice	but	from
7	because	want	will	have	learn	what	but	society	learned	practice
8	For	and	know	improve	their	out	can	experience	and	can
9	Only	society	more	master	many	just		you	can't	that
10	Our	hard	society	with	real			knowledge	will	will
11	Think	not	have	study	and		you	have	you	world
12	practice	but	also	enrich	English		what		world	not

词语型式统计对研究分析某一搜索词在一定跨距内的词语搭配和用词格局具有重要意义，它不仅为说明词语用法和比较同义词群提供必要的统计信息，还能提示围绕某一词语在各个词位上的词语聚合关系。

3.4　词丛统计(word cluster)

词丛统计对预定长度的词语组合在语料库中全程查找，并计算其复现频数。其统计结果是各种长度的词丛表。在 *Wordsmith Tools* 中，词丛统计围绕搜索词计算预定长度的词丛和频率。如搜

索词 *experience* 在 BROWN 语料库中的频率最高三词词丛有 *the experience of*、*the experience in*、*a lot of*(*knowledge*)、*knowledge and experience*。在 TACT 中,词丛统计的方法是对语料库全程一次性计算,把所有预定长度范围内各种长度的词丛统计出来。如使用者可以预定长度范围为从两个词到六个词,TACT 通过计算,生成一个从两词到六词的词丛总表,等长词丛表内的词丛以字母或频率高低排列。词丛统计可以验证词组、短语以及搭配在某一语料库中的分布和典型特征。

3.5 主题词提取(key word list)与词图(plot)

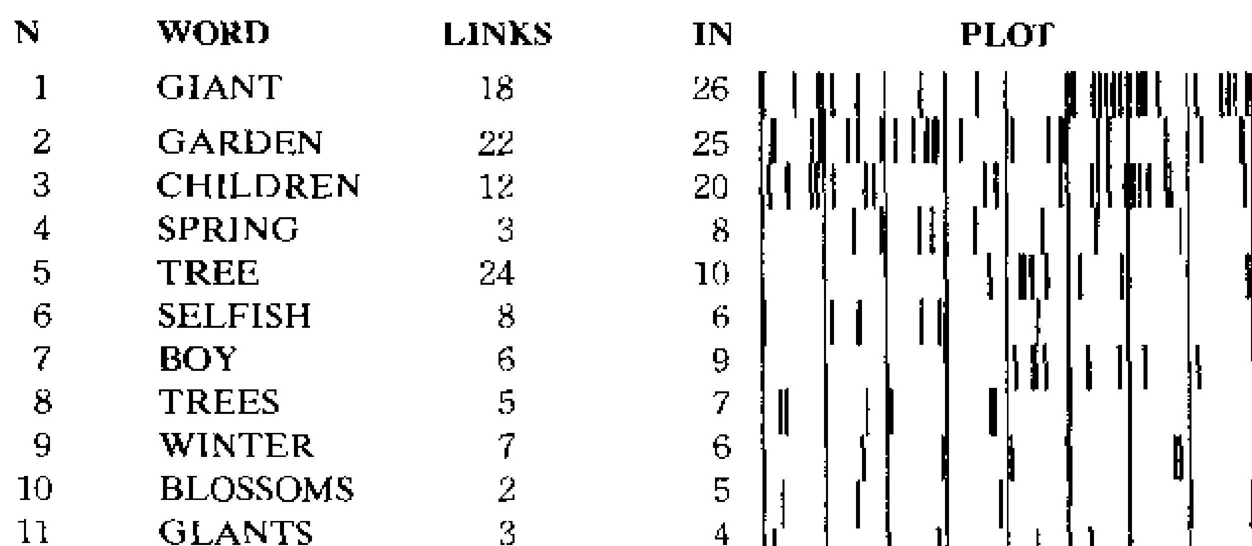
主题词提取和词图功能是 *Wordsmith Tools* 所独有的。上述各种功能都是围绕单一词语(搜索词)在短语和句子层面的用法分析,但无法观察和分析整个语篇中的词语关系。在不同的语篇中,词语的选择和分布存在差异。不同主题的文本通过词语的选择体现出来。另外,正如句子或短语中的词语具有搭配和语法关系一样,语篇中的词语也具有共现关系。通过提取和分析语篇中具有超常频率的词以及具有共现关系的词语或词群,可以确定语篇的主题和表达该主题的词集,进而研究作者对某一主题的心理词符和知识表述问题。就学习者语料库而言,通过主题词研究,可以观察学习者对某一主题所使用的词语以及词语之间的关系。主题词提取的方法首先是对比某一完整连续文本和一个更大的参照语料库(reference corpus),把语篇中差异显著的词语提取出来,生成为一个主题词表。差异的显著性通过 X^2 检验用 P 值表示($P \leqslant 0.000\,01$)。词图统计是根据主题词表,计算出各个主题词在语篇中的位置分布。词图的意义主要在与对某一连续文本的词语分布进行统计和计算,如果是像 JDEST、BROWN、LOB 或 CLEC 这样的杂合性语料,尽管 Wordsmith Tools 也能统计出单个词的词图,但由于缺乏相对应的文本信息作为参照,其价值有限。另外,单个词图只有同其他词图放在一起进行比较,才显示出真正意义。对

搜索词进行各种词图统计,前提必须是所索引的文本是一连续的整体,如一篇小说或一篇论文。通过词图统计,可以观察主题词在语篇中出现的先后顺序和分布密度,直观地分析该语篇的主题发展或情节的推进与词语的关系。如利用该方法对王尔德(O. Wilde)的《自私的巨人》(*The Selfish Giant*)进行分析,提取出主题词并列表(参照语料库为王尔德的20多部作品)。从下表可以看出,排在前几个的主题词依次为 *giant*、*garden*、*children*、*spring*、*tree*、*selfish*、*boy*、*trees*、*winter*、*blossom* 等。这些词反映了这一作品的主要主题信息:主要人物为 *giant*、*children*、*boy*;时间为 *spring*、*winter*;背景是 *garden*、*grass*、*frost*;关键描述信息为 *selfish*、*blossom*。

表 5.4 *The Selfish Giant* 的主题词表(部分)

N	Word	Freq.	Oscar7.txt %	Freq.	Oscar.txt %	Keyness	P Value
1	Giant	26	1.56	55	0.05	118.3	0.000 000
2	Garden	25	1.50	100	0.09	87.8	0.000 000
3	Children	20	1.20	92	0.09	65.5	0.000 000
4	Spring	8	0.48	28	0.03	29.8	0.000 000
5	Tree	10	0.60	66	0.06	26.7	0.000 000
6	Selfish	6	0.36	14	0.01	26.3	0.000 000
7	Boy	9	0.54	64	0.06	22.9	0.000 002
8	Trees	7	0.42	36	0.03	21.6	0.000 003
9	Winter	6	0.36	28	0.03	19.5	0.000 010
10	Blossoms	5	0.30	17	0.02	18.9	0.000 014
11	Giant's	4	0.24	8		18.5	0.000 017
12	Hail	4	0.24	8		18.5	0.000 017
13	Grass	6	0.36	36	0.03	17.0	0.000 038
14	North	4	0.24	11	0.01	16.5	0.000 049
15	Frost	4	0.24	13	0.01	15.4	0.000 088

再通过以下词图观察主题词在文本中的位置分布。主角 *giant* 一词在全篇都得到使用，故事的情节基本都是围绕他展开的；但在故事的后半部分，该词的使用密度突然增大，显示了故事的冲突发展高潮阶段。自私的巨人把孩子们从花园赶跑，春光明媚的花园突然转为凛冽的寒冬，故事线索集中在 *giant* 一人身上。与之相对应的是，另外的主角 *children* 一词在故事的开始部分分布均匀，而在故事中部出现空白，在故事结尾处出现密集分布。这表明了孩子们被巨人赶出花园后，重新回到花园玩耍的情节发展。*boy* 一词稍后出现，显示故事冲突解决的线索安排位置。表示象征意义的词 *blossom* 只在故事开头和结尾处出现，显示了故事优美的开始和完满的结局。



N	WORD	LINKS	IN	PLOT
1	GIANT	18	26	
2	GARDEN	22	25	
3	CHILDREN	12	20	
4	SPRING	3	8	
5	TREE	24	10	
6	SELFISH	8	6	
7	BOY	6	9	
8	TREES	5	7	
9	WINTER	7	6	
10	BLOSSOMS	2	5	
11	GIANTS	3	4	

图 5.4　*The Selfish Giant* 主题词词图统计

4　个案分析实例：*obtain*、*knowledge*、*success* 在不同语料库的用法分析

亨斯顿(S. Hunston)和弗朗西斯(G. Francis)(1998)把动词的型式分为以下几种：1)词项单独出现：V。2)词项后跟某一词类：

V + n; V + adj; V + adv。3)动词后跟从句或动词的其他词形:V + that; V + wh; V + wh + to-inf; V + to-inf; V + -ing; V + inf; V + -ed。4)动词后跟介词短语: V + about + n; V + as + n; V + at + n; V + prep。5)动词后跟名词词组加其他词类:V + n + n; V + n + adj; V + n + adv。6)动词后跟名词加从句或动词的其他词形: V + n + that; V + n + wh; V + n + wh + to-inf; V + n + to-inf; V + n + -ing; V + n + inf; V + n + -ed。7)动词后跟名词加介词短语:V + n + about + n; V + n + as + n; V + n + at + n; V + n + prep。8)动词前有 it 为前导词:it + V + that; V + it + adj + that。9)动词前用 there 作主语:there V + n; there + V prep/adv (Houston and Francis 1998)。下面对动词 *obtain* 在学习者语料库和英语本族语语料库在型式,即搭配类联接的差异作一对比: *

表 5.5 *obtain* 一词搭配类联接的对比数据.

	CLEC	LOB	BROWN	JDEST
V + n	+	+	+	+
be + V-ed + as + n	–	+	+	+
be + V-ed + prep	+	+	+	+
V + for + n + n	–	+	+	+
V + n + prep + n	+	+	+	+
V + n + to-inf	+	+	+	+
be + V-ed + ing		+	+	+
V + that	+	–	–	–
V + adj	+	–	–	–

* 表中 + 号表示该型式在库中出现, – 号表示未出现。

从上表中可以看出，学习者语料库中动词 *obtain* 的词类类联接与本族语者相比，类联接的变化要少，对一些较复杂的类联接如 be + V + -ed + as + n，V + for + n + n，be + V-ed + -ing 等较少用到，表明学生对该动词的用法尚未完全掌握。另外，学习者语料库中还有两个异常类联接，即 *obtain* 后跟从句(V + that)以及 *obtain* 与形容词的组合。这两种型式在英语本族语语料库中一次也没出现。

关于类联接的对比为我们了解学习者的中间语提供了有价值的信息，但还不能使我们深入了解某一类联接在学习者语言运用中的行为实质。如果把有关类联接的信息与搭配结合在一起，就可以得到更有价值的信息。在下表中，通过查询"obtain"一词在四个语料库中的使用实例，调查它与哪些词经常搭配，并对比其语义限制在不同的库中有什么差异。限于篇幅，下表仅列出与该词最常搭配的名词(按频率列出前 10 个词，在左右 ± 4 个词之间的非连续性搭配)：

表 5.6　搜索词 *obtain* 的搭配词对比

CLEC	LOB，BROWN，JDEST
water，knowledge，money，experience，success，plenty，time，education，wealth，interest	results，values，experiment，data，information，difference，solution，possibility，item，time

可以看出，学习者语料库中 *obtain* 一词所吸引的名词与本族语者语料库中该词所吸引的名词差异很大。在 LOB、BROWN、JDEST 三个语料库中，*results*、*value*(*s*) 和 *experiment* 出现在动词 *obtain* 周围的频率非常一致，表明这三个名词与所查动词的搭配强度(collocability)相当大。但学习者语料库中却未出现这种搭配，而学习者语料库中所出现的搭配词(collocates)在本族语语

料库中也很少出现。与前表中的类联接的差异情况相比，可以看出学生对该动词的语法规则的掌握比较接近英语本族人的语言运用，但对该动词的语义选择限制却呈现出很大的差异。然而，应该指出的是，这种对比只能提供一种直观粗略的描述，信息提取和呈现的方法存在严重缺陷。首先，各个语料库的容量有差异(LOB 和 BROWN 分别为 100 万，JDEST 为 400 万，CLEC 子语料库为 50 多万)，受其容量的影响，某一搭配词是否出现以及出现的频率代表不同的意义。其次，CLEC 子语料库是命题作文，词语的使用受到题材限制。再次，由于对搭配词的索引查询(concordancing)是以某中心词左右四个词为幅度进行的，我们虽然能得到搭配词的搭配频率，却很难了解这些搭配词与中心词的确切关系。为了深入了解某一搭配的语义关系，就需要对所索引的词进行限制，如主谓关系、描述关系、动宾关系等。这样，通过查询，我们就可以得知某一词项的搭配的语法和语义限制关系。最后，通过对比同类搭配的聚合词群，进而分析该聚合词在语义和类联接的内在关系。下面以学习者语料库中出现频率较高的一组词(*knowledge*、*ability* (*-ies*)、*success*)为中心词，以 V + N/ING(noun group)为关系限制，在各个语料库中查询，限定跨距为中心词左边 5 个词，在查询中，对动词的各种形式作原形还原处理，如 *learn* 包括 *learns*、*learned*、*learning*、*learnt* 诸形式。查询结果见表 5.7、5.8、5.9。

从表中可以看出，学习者语料库中与 *knowledge* 搭配的动词(与所查名词为动宾关系)按频率依次为 *learn*，*get*，*have*，*study* 等。除了 *have* 一词与 *knowledge* 的搭配在四个语料库中都有重复出现频率外，*increase* 与 *knowledge* 的搭配只在 CLEC，LOB，和 BROWN 三个库中出现。如把搭配词的对比作一划分，可分为以下几种情况：1)同一搭配在四个库中都有不同频率的复现，表示学习者对目的语已经掌握的部分，如：*have* (*knowledge*、*ability* (*-ies*)、*success*)；*show* (*ability*)；*make* (*success*)。2)学习者语料库中出现的搭配只在本族语语料库中有低频率的出现，如 *accumulate*、*acquire* (*knowledge*)；*improve* (*ability*)；*achieve*

(*success*)等。3)搭配只在学习者语料库中出现,而本族语者语料库中却从未出现,如 *study*、*master*、*grasp*、*enrich*、*enwide*、*practice*(*knowledge*);*find*、*promote*、*prove*、*raise*、*train*、*use*、*learn*(*ability*);*get*、*want*、*gain*、*reach*(*success*)。4)在学习者语料库中出现的搭配在本族语语料库中频率低于 2 次,搭配关系不显著,如 *learn*、*get*、*increase*、*obtain*、*enlarge*、*broaden*(*knowledge*);*display*、*practice*(*ability*);*win*(*success*)。5)搭配在四个库中频率都低,基本排除搭配的可能性,如 *widen*、*deepen*(*knowledge*)。值得注意的是,学习者语料库的容量小于其他三个库,某个搭配的原始频率所代表的意义是不一样的,如果对四个库都按 100 万词的容量进行标准化处理,则每个搭配的频率比值差异会更大。

Sample concordance lines for *study* * + *knowledge* in the written corpus of CLEC

10	according to some things. So I will	study the knowledge of society before g
11	l young, we have the best energe to	study new knowledge and master it very
12	ith it in order that consumers can	study more knowlecge about the commoditi
13	ities productor . Second, we should	study the knowledge about commodities.
14	a high mark in examation, but also	study new knowledge well. In a word, pr
15	g job may give them many chance to	study more knowledge and accumulate much
16	to adapt new surroundings and to	study the knowledge about his new job.
17	For example, a person wanted to	study the knowledge of accounting to lo
18	y; secondly, it affort a chance to	study new knewledge and broad their eye
19	improve themselves, and make them	study much knowlecge, and the money is a
20	f our college life. Do our best to	study more knowledge. But some students
21	think that is interesting and can	study much knowlecge. The second view is

表 5.7 "V + N"类联接中 *knowledge* 的动词搭配

	CLEC	LOB	BROWN	JDEST
LEARN	246	0	0	1
GET	51	0	0	1
HAVE	41	14	9	19
STUDY	37	0	0	0
MASTER	22	0	0	0
GRASP	19	0	0	0
ACQUIRE	9	0	0	10
INCREASE	9	1	1	0
ENRICH	8	0	0	0
OBTAIN	8	0	0	2
ENLARGE	8	1	0	0
BROADEN	5	0	1	0
ENWIDE *	5	0	0	0
PRACTICE	4	0	0	0
WIDEN	1	1	0	1
DEEPEN	1	0	0	2
ACCUMULATE	1	2	0	2

表 5.8 "V + N"类联接中 *ability* 的动词搭配对比

	CLEC	LOB	BROWN	JDEST
IMPROVE	54	0	0	6
HAVE	40	4	7	26
SHOW	15	1	2	2
DEVELOP	14	0	0	3
FIND	7	0	0	0
PROMOTE	5	0	0	0
PROVE	9	0	0	0
DISPLAY	3	1	1	0
RAISE	5	0	0	0
TRAIN	4	0	0	0
USE	5	0	0	0
PRACTICE(SE)	3	1	0	0
LEARN	4	0	0	0
ENHANCE	1	0	0	7

表 5.9 "V + N"类联接中 *success* 的动词搭配对比

	CLEC	LOB	BROWN	JDEST
MAKE	18	3	3	1
GET	26	0	0	0
WANT	6	0	0	0
GAIN	18	0	0	0
LEAD	16 *	1	0	0
ACHIEVE	14	1	1	11
HAVE	3	6	4	13
WIN	4	1	0	0
REACH	2	0	0	0
OBTAIN	5	0	0	2

* Two cases, *lead us to success* and *lead me to success*, are included.

以上列表对比了学习者的常用搭配与英语本族语者的语言运用，所得数据表明，差异是非常明显的。

通过以上对比分析，可以看出，中国大学英语学习者在动名搭配的使用中语法准确性远远高于语义语用上的合适性。从目的语的标准来看，某些名词所吸引的动词聚合不符合英语本族语者的语言直觉。正是这种差异造成了学习者的中间语与目的语之间真正的距离。正确选择搭配中所吸引的词对本族语人来说是自动的，直觉的，且选择范围非常有限。对于外语学习者来说，这种选择的范围要大得多，他们会认为很多聚合词在语义和句法上都是适当的。本森(Benson)(1990)曾经针对搭配设计了一套填空测试，让英语本族人选择合适的搭配词(见下表)。斯麦迦(Smadja)(1993)认为，这些句子中的搭配词对本族语者是很容易填出的，但对第二语言学习者来说就很难自如地发现适当的词，他们会觉得右栏中很多词在句法和语义上都是合适的：

表 5.10 填空练习

sentence	candidates
"If a fire breaks out, the alarm will ______."	"ring, go off, sound, start"
"The boy doesn't know how to ___ his bicycle."	"drive, ride, conduct"
"The American congress can _ a presidential veto."	"ban/cancel/delete/reject"
"Before eating your bag of microwavable popcorn, you have to _ it.	"cook/nuke/broil/fry/bake"

从错误分析的角度来看，学习者在搭配运用上的特征大多数来自母语的影响，可以通过语义解释把学习者的句子翻译为母语，如下列

句子：

sample lines from CLEC	*Possible translations in Chinese*
So we must do more practice to increase our knowledge.	所以我们应更多地通过实践来增长知识。
Then we can grasp the knowledge well and can use it freely.	这样我们就能更好地掌握知识并能自由地运用它。
He must not only study the book knowledge, but also study the social knowledge.	他不仅要学习书本知识，还要学习社会知识。

在汉语中，"知识"一词与"增长"、"掌握"、"学习"等词的搭配都是正常的，符合我们对母语的直觉。但在英语中，类似的搭配并不常见。在这里，母语的影响显而易见。应用语言学家的研究表明，语言迁移在外语学习中的影响取决于以下几个因素：1）成年的学习者比儿童更容易发生语言迁移，且年龄越大，迁移的影响越显著；2）正式的语言学习比非正式的语言习得更可能发生语言迁移；3）语言迁移受学习动机和目的的影响(Corder 1981)。大学英语的学习者属于成年学习者，学习环境是正式的语言教学，所以以语言迁移来解释其语言运用，可以使我们更深刻地了解学习者的语言发展过程，以及在此过程中母语与目的语相互作用所产生的影响。当然，仅以语言迁移难以解释所有的中间语特征。就迁移而论，语言迁移也不是唯一的因素。塞林克(L. Selinker)(1972)认为，除了语言迁移，教学迁移(即学习者受语言教师语言的影响)、交际迁移(即学习者受交际事件的影响)，以及策略迁移同样会构成中间语的基本特征。在基于语料库的研究中，语料库的容量、语料源在词汇数量和结构上的差异，都会影响某项搭配在库中的行为。

5 可用资源与索引软件

使用语料库及索引软件要具备的条件是：一台计算机、语料库（以纯文本形式存储的电子文本）、语料库索引软件。语料库文本可以通过网上下载或编制自己的专用语料库。索引软件可以从相关网址获得。下表中简要列举了可获得的电子文本和语料库资源及网址（参见 Ball 1997）：

资源名称	获得方法	说明
语料库列表	（1）发电子邮件：*MAJORDOMO@UIB.NO* 正文处填写：subscribe corpora；（2）网址：*http://www.hd.uib.no/corpora/archive.html*	可获得各种语料库资源的详细列表。
“语言研究用电子语料库及相关资源通览”（Jane Edwards 1993）	通过 FTP 登录：cogsci.Berkeley.edu/pub/CorpusSurvey.ascii	详尽介绍了各种语料库及电子文本资源的位置和获得方法。
Georgetown 电子文本工程通目（CPET）	GOPHER 至：gopher.georgetown.edu	
牛津文库（OTA）	*http://sable.ox.ac.uk/ota/*	
语言学数据集团（LDC）	*http://www.ldc.upenn.edu/*	
Project Gutenberg	*http://www.promo.net/pg/*	
BNC Online	*http://thetis.bl.uk*	英国国家语料库容量为 1 亿词，通过网上检索，每次可浏览 50 行索引。另外，也可下载专用软件 sara，注册后限期使用。
Michael Barlow 博士语料库语言学主页	*http://www.ruf.rice.edu/~barlow/corpus.html*	语料库语言学主页。详尽介绍了各种语料库及应用软件。

在选择外语课堂教学所用的语料库索引软件时，我们应从以下几个方面考虑：1）易操作性：包括用户界面是否友好，运行环境是否适用常用的平台，如 WINDOWS 操作系统，指令是否复杂，以及是否提供实时帮助等。目前网上可下载的索引软件各有不同的操作系统，如 DOS、WINDOWS、SOLARIS、UNIX 等。在下载前一定要弄清楚该软件的平台要求。2）文本预处理要求，即语料库文本在进入索引程序前是否需要生成特定的格式或标识。如 TACT 索引软件在处理文本前要求将文本生成一个标注文件（Mark-up file），而标注文件中信息需要使用者自己填写。这种软件对自己动手建立的语料库很有用处。3）检索结果展示直观性以及是否允许在屏编辑。通过在屏编辑，教师可迅速提取所需信息，减少中间环节。4）功能丰富性。一般来说，功能的丰富与易操作是一对矛盾。对于外语教学来说，易操作是最重要的。教师可能对那些功能丰富但操作复杂的软件失去耐心，兴趣索然。初学者可先从功能简单、容易操作的索引软件入手，如 MicroConcord，然后再使用功能较复杂的索引软件。5）开放性，即每次处理语料量是否有限制。例如，一个 100 万词语料库如果必须分 4 次才能检索完，对其结果还要进行再次人工统计，使用者就会不胜其烦。6）自由软件与技术支持。即该软件是否能免费获得，以及不断得到升级支持。索引软件一般分为以下几种：（1）自由软件。自由软件可以从网上免费下载，个别软件要求注册或给作者发一封电子邮件，如 MicroConcord、Tact、Xtract、Paracorp 等。（2）演示版。索引软件的演示版本身包含完整的程序和功能，但需要购买注册后解密才能使用。否则，该软件各种功能的使用非常有限，如 Wordsmith Tools。（3）限期使用的软件。这种软件从网上下载后可获得一个注册号，该注册号在使用一定次数后即失效。如 BNC 的 Sara。（4）商业软件。需要购买才能使用，如 Wordsmith Tools。以下对几个常用的自由软件的特点、使用方法、以及获得途径作简短介绍和评述，其他软件参见附录：

MicroConcord

该软件基本特点是:用户界面简单,展示清晰;结果快速,即时检索,无须对文本进行预处理,但处理文本量有限。

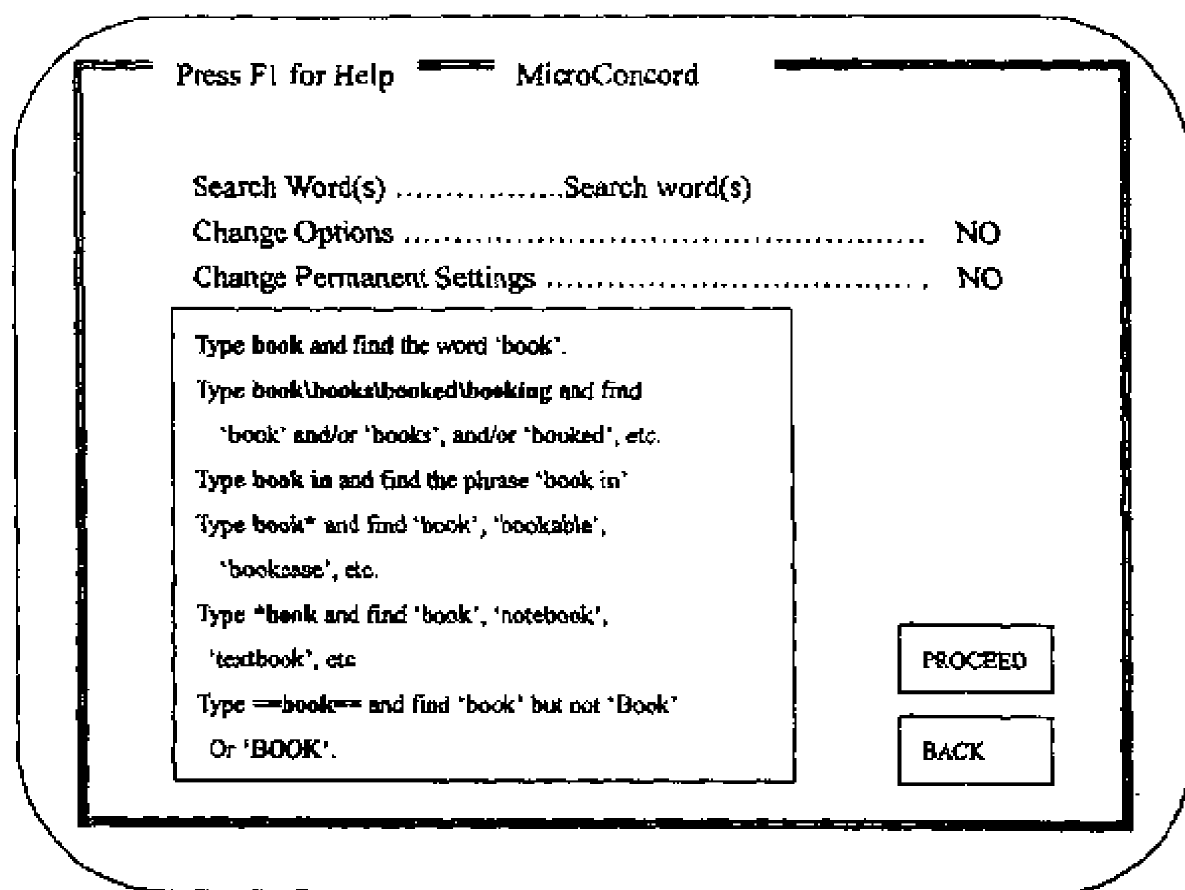
希金斯(Higgins)(1991)认为 MicroConcord 是"课堂索引器"。课堂索引器有两大功能:1)通过检索功能词以使学生自己发现语法规则;2)检索同义词对,以使学习者区分其用法。课堂检索程序要能提供快速的检索结果和清晰的展示(Higgins 1991:92)。但是其功能简单也同时是它的弱点,如不能生成词表及有关统计信息,在查询中难以对语料的各种变量进行控制。另外,由于它只使用 DOS 的常规内存,所以索引的数量有限,对于大型语料库来说已经力不从心。当然,使用者可以对语料分批进行处理,然后再累积计算,但工作量会过于浩繁。尽管如此,Microconcord 仍不失为一种首选的语料库入门用的索引工具。

MicroConcord 界面简单直观。由用户指定搜索词、搜索文件所在目录和文件名。允许同时指定多达 900 多个目录下 500 多个文件的实时搜索。搜索过程中实时显示结果和所在文件及搜索词的分布信息。显示窗口展示 KWIC(key words in context),并动态显示光标所在行的文件名和位置。使用者可以编辑显示结果、重新排序、观察关键词所在语篇,获得左右搭配词表,并把结果存为一个单独的文件。整个过程都提供帮助提示信息。该软件无须安装,使用者只需把下载的文件拷入硬盘,即可使用。使用 MicroConcord 的基本步骤如下:

1. 下载 MicroConcord。该软件有两个版本,更新版 Overlay version 可从 *http://grwy.online.ha.cn/liwenzhong* 下载。
2. 把准备好的语料库文本文件放在一个固定的目录下,虽然 MicroConcord 支持多目录检索,但把检索文件集中放置更便于管理。语料库文本可以是未经赋码的生语料,也可以是经过自动赋码或手工赋码的语料。
3. 把 MicroConcord 执行程序及相关文件拷入硬盘,存放在同一

目录下。

4. 执行 MicroConcord 程序。在 DOS 状态下，进入该程序所在目录，键入 Mconcord，按回车键。界面如图所示：



5. 进入程序后，根据提示，键入要检索的关键词。可以同时检索多个关键词或同一关键词的各种形式，词与词之间用反斜杠隔开，如：*study* \ *work* \ *learn* \ *read* \ 或 *study* \ *studies* \ *studying*。也可以使用统配符“ * ”，如 *stud* * 可以检索出 *study*、*studied*、*student*、*studying*。

6. 改变文件缺省路径及目录，并指定要检索的文件类型。如“C:\CORPUS”指定要检索的文件的位置在 C 盘 CORPUS 目录下；“ * .TXT”则指定该目录下所有以“TXT”为后缀的文件；“ * . * ”指定该目录下所有类型的文件。

7. 键入关键词前后的搭配词。此项是可选项。
8. 指定关键词前后搭配词检索跨距(Span)。如"3,3"表示命令程序检索关键词左右3个词内所指定的搭配词,并把结果显示出来。

按回车键让程序运行。在显示窗口可以通过键盘操作调整和编辑检索结果。如按"Ctrl+方向键"可以依次对关键词、搭配词按字母或频数排序;按回车键(Enter)可以显示某一行的整篇文本;按空格键可以把关键词隐去,编制成填空练习;按"C"键可以观察搭配词表,并可对词表进行排序;按"S"键可以存储结果。保存的文件可以在MS WORD或WPS2000中打开进行处理和编辑。下载网址:*http://www.liv.ac.uk/~ms2928/homepage.html*;国内下载:*ftp://ling.gdufs.edu.cn/gui*, *http://grwy.online.ha.cn/liwenzhong*。

TACT

TACT(Text Analysis Computing Tools)是加拿大多伦多大学开发的语料库索引软件包。它包括16个相对独立的执行程序,具有全文检索、语境中的关键词索引(KWIC)、词表生成与词频统计、搭配词自动提取、语料比较等强大功能。该软件包原来用于文学作品研究,对研究作家的风格、用词等提供方便的查询。该软件包由多伦多大学的布拉德利(J. Bradley)、普拉苏提(L. Presutti)、斯德尔斯(M. Stairs)和兰克夏(*I. Lancashire*)在1984年联合开发完成,如今作为共享软件允许使用者从网上免费下载(共两兆多字节,需要注册),国内外语教师也可以从广东外语外贸大学FTP服务器桂诗春教授目录下匿名登录下载(*ftp://ling.gdufs.edu.cn/gui*),或在多伦多大学主页下载:*ftp://chass.utoronto.ca/pub/cch/tact/tact2.1/*。与MicroConcord一样,该软件也是基于DOS平台。不同的是,需要把语料库文件做成TACT可处理的库文件,并进行标注处理,如对语料的分类、作者以及其他信息

插入控制码，并在 TACT 上另外做成一个标注文件。TACT 允许对各种变量进行控制索引，并能提供某一索引词或特征在各个变量中的分布。它可以满足不同的研究需要，按照要求计算特定的数据信息。但是要真正掌握并熟练运用它则需花费时日。另外，其处理语料有限，最大容量可处理 1 兆字节的文本文件（做成库文件后则会变成 5 兆左右）。如果语料文件太大，则会导致系统崩溃。如对大型语料库文件进行处理，就需要把文件分批处理，操作较为复杂。但是，TACT 是目前国内外语教师能免费得到的功能最全的语料库索引软件。值得一提的是，近两年由布拉德雷和洛克维尔（G. Rockwell）开发的 TactWeb1.0 是基于 TACT 2.1 开发的网络索引软件。该软件可以作为插件放置在主页上与因特网或局域网服务器上的 TACT 的 TDB 文件连接起来，允许网络用户在线索引。该软件下载网址为：*ftp://ic-unix.ic.utoronto.ca/pub/tactweb/readme.html*。

Wordsmith Tools

Wordsmith Tools 是利物浦大学斯科特（M. Scott）在 MicroConcord 基础上重新编写的最新的语料库索引工具。该工具在 1996 年公布于众。Wordsmith Tools 在保留了 MicroConcord 所有功能以外，还增加了以下新的功能：1）生成词表，可按词频、字母顺序分别排列，并提供各种统计信息。2）关键词提取。通过小语料库（如 100 万词）与另一大语料库（如 5000 万词或更大）的词表词频比较，可统计出小语料库中超常词频词，构成该库的关键词词表。进一步统计分析关键词在单个语篇中的分布，再生成该语料库的主要关键词（key key words）（Scott 1997）。在主要关键词表的基础上，还可查出某一关键词的联想词汇（associates）。此项功能对研究同题材语料的主题意义和论题，以及词汇之间的意义关系提供方便的统计手段，代表了目前语料库技术的前沿水平。3）在索引查询中，同时提供词汇词语型式（Lexical patterns）表和搭配词位置分

布等极有价值的信息,使研究者可以从多种角度对词汇运用进行分析。4)查询结果可以很方便地转换成表格格式,并读入到像 MS Access、Excel 等数据库中进行进一步分析统计。值得一提的是,Wordsmith Tools 是基于 Windows 平台的索引软件,可在 Windows 95 或 Windows 98 中使用,其处理语料的数量仅受计算机内存和硬盘大小的限制,也就是说,它一次处理的语料量几乎是无限的。另外,它操作简便,容易使用,风格与 Windows 保持一致。但是,该软件需要先从牛津大学出版社网页上下载,然后再向其购买授权证书和密码(个人用户约 60 英镑)。如不购买,则只能使用其演示版,不能实用。下载网址:*http://www.oup.co.uk*。

TEC Concordancing Tools

TEC 是 *Translational English Corpus* 的首字母缩写。它是英国曼彻斯特大学理工学院(UMIST)语言工程系翻译研究中心开发并维护的当代翻译英语语料库。该语料库由各种语源译成英语的语料构成,主要分为三种文类:传记、文学作品与报纸新闻。该库主页网址为 *http://ubatuba.ccl.umist.ac.uk/tec/*。对该语料库的索引可通过两种方式进行。一是浏览主页并通过索引插件获得索引;再就是下载独立运行的客户端软件 TecCon1.1,安装在个人 PC 机上后,通过上网远程访问 TEC 语料库,获得索引结果。下载网址为 *http://ubatuba.ccl.umist.ac.uk/tec/download*。该索引软件提供词语索引行与搭配词表,以及各词语索引行的扩展语境。允许用户把索引结果存储在自己的硬盘上。该软件及语料库是免费的,但由于版权问题,TEC 语料库不能下载,且只能使用该软件对 TEC 进行索引。

Concordance

Concordance 是瓦特在 1999 年开发的语料库索引软件。该

软件除了一般文本索引软件所具有的功能外，其独特之处是能够把索引结果自动生成 HTML 网页，供在线浏览。该软件为独立软件，可利用它对任何语料库文本进行索引分析。*Concordance* 属于共享软件，下载后可试用 30 天。购买方式可以通过在线注册、电汇，或直接汇款给 R. J. C. Watt 本人。必须先下载软件，注册后可获得注册码和密码。下载网址：*http://www.rjcw.freeserve.co.uk/dl.htm*。作者 Email 地址：*R.J.C.Watt@dundee.ac.uk*。

除上述索引工具外，其他索引工具还有英国伯明翰大学 COBUILD 语料库索引器、用于 BNC 语料库的 SARA、专门用于搭配研究的 X-tract；以及独立索引软件 Word Cruncher、Lexas；用于 Mac 平台的 MonoConcord 等，不再一一详述。另外，上海交通大学语料库索引器(JDEST)，以及广东外语外贸大学开发的索引器也值得一用。但是 JDEST 语料库索引器是针对交大科技英语语料库开发研制的，难以移植，这里不再一一介绍。

6　小结

语料库索引在外语教学和学习中的应用及意义包括以下几个方面：1）基于语料库索引的 DDL(数据驱动学习，Data-driven Learning)。由于语料库索引提供搜索词用法的真实语境、词汇搭配以及频率信息，通过词语索引行可以开发出实时词汇练习、同义词比较、搭配词组练习等。学习者可以通过所提供的语境选择词汇。目前可以通过三种基本手段实现：其一是开发出独立的 DDL 软件，把语料库词语索引行以及词汇练习一同打包。使用者通过程序界面选择调用要进行练习的项目。比较成功的 DDL 软件如 Tim Johns 开发的 Xcontext 就属于此类。这种软件的优点在于携带方便，对硬件要求低，适合个人在 PC 机上学习。但软件封装后，

就成了一个封闭系统，难以对语料更新，且提供的词语有限。其二是与其他学习材料结合起来，针对语篇中词汇和搭配制作基于语料库索引的交互式练习，再结合动态超文本(DHTML)格式转换为可在网上传递的课件，供远程课堂或局域网上网络教室使用。这种方法的优点在于充分利用网络优势，集成超媒体技术手段，实现各种媒体在单一平台播放；可以随时更新和修改课件内容，并与互联网资源动态连接。词语索引练习具有交互性，可提供即时反馈。其三是利用语料库索引进行课堂实时演示，通过教师的参与和指导进行语言学习。这种方法的特点是灵活开放，探索性强。但要求具备可用语料库以及语料库索引软件，还需要对操作者进行必要的培训。2)利用语料库索引开发网页索引。这种技术把单个语料库所有词语索引转换为超文本格式的框架网页。使用者可通过字母或词表索引观察词语索引行和语境信息，如作者利用 Webconcordance 制作的“大学英语学习者语料库在线索引”。这种方法可以实现语料库资源共享，使那些不具备语料库其工具的远程用户或局域网教室使用，操作简单方便，不需要特殊培训。3)语料库在线索引。通过前端语料库索引插件与放置在服务器上的数据库相连，实现网络用户对语料库的直接访问和索引。如 TACTWEB 允许用户制作自己的语料库数据库文件，并使用 TACT 插件实现语料库资源共享。

把语料库索引与多媒体课件开发有机地结合起来，能够充分发挥 DDL 的探索性和课堂教学的针对性。语料库资源的共享为开展交际教学和个人化学习，解决教学材料的真实性问题，缓解我国外语教学中教师语用知识和语言直觉欠缺这一瓶颈问题，提供了一个新的路径。同时，语料库在外语教学中的应用也要求外语教学人员在教学思想和方法上的革新和转变。



第六章　英语词语搭配的种类

引　　言

词语搭配是词与词的有限组合。与自由组合的情况不同，构成有限组合词的搭配力受诸多因素的制约，只能和有限范围的词结合在一起使用。然而，词语搭配又和成语的情况不同。成语基本上是一种固定词组；在成语里，词的搭配能力已被穷尽，已无搭配范围可言。但是，构成有限组合的词却有一定的自由度去和一定范围内的词结合为搭配伙伴，一起使用。“搭配”一词的涵义就在于它揭示了词语组合行为的有限性这一特点。然而，制约词语组合行为的因素是十分复杂的。由这些复杂的因素导致而形成的搭配又各有特点，可分为几大类别。关于搭配的类别，弗斯(Firth 1957)曾经谈到一般搭配、个人搭配、技术性搭配等情况，但没有进行过界定和探讨。在本章里，我们试图根据词语搭配的使用领域(areas of use)，对搭配力有重要制约作用的词的义项(sense)及其特点，以及搭配的因循性程度，探讨词语搭配的种类及其有关特点。词语搭配将被分为一般搭配(general collocations, usual collocations)、修辞性搭配(figurative collocations)、专业性搭配(specialized collocations)和惯例化搭配(institutionalized collocations)逐一加以讨论。

1　一般搭配

一般搭配是一种在普通语言里使用频率极高的习惯性词语组

合方式,而且在不同程度上被用于各个专业领域。一般搭配的构成大致有三种情形。其一是关键词(key word),或者说节点词(node word),和意义相关的一个有限范畴的词项相互搭配,表达一定的意义;其二是节点词和几个意义不同的词项相互搭配,表达不同的意义;其三是节点词和某个特定的词项结合,构成一种在形式上近乎于成语的搭配。

1.1 节点词和意义相关的有限范畴的词项搭配

在这一类情况中,属于有限范畴的词项在意义上是相关联的。这些词要么都有一种相同或相似的某种内涵意义。要么是属于同一个语义范畴(semantic category)的词,所指的事物类似或相近。下面是从5000万词的COBUILD语料库中提取的关于"commit"一词的词语索引:

1	Merely by staying on, did not	commit a criminal offence. [p] Both the
2	she felt she would never be able to	commit a serious sin again. Thirty-one
3	a knock-out or somebody's going to	commit a major faux pas, and outside of
4	prison walls; or for governments to	commit abuses that won't cost their
5	Exodus 20 14), 'Thou shalt not	commit adultery" has become 'Be faithful
6	for a crime they knew he did not	commit and you know Monroe County is like
7	of an Account Holder or if you	commit any breach of the Conditions,
8	for their own uses. They will	commit any crime, but never in passion.
9	Months inside for a crime he didn't	commit. But in the meantime he had found
10	countries (and therefore free to	commit crimes there) that country's right
11	f Mold, Clwyd, denies incitement to	commit grevious bodily harm. The case
12	an action towards you cause you to	commit murder?" [p] Sommers turned from
13	the activist students, he saw them	commit no vandalism; on the contrary they
14	had ended my life. Jane I would	commit suicide. I wouldn't want someone
15	claimed not only that Green did not	commit the crime but that the body was not

由上述词语索引可以看出，*commit* 一词的主要右搭配词包括 *crime*、*offense*、*suicide*、*murder*、*abuse*、*sin*、*harm*、*vandalism*、*breach*、*adultery* 等。在这些搭配词中，*suicide* 和 *crime* 出现的频率最高。如果用统计手段 T-score 来测量各搭配词的显著性程度，则可得到下述信息：

表 6.1　*commit* 一词的右名词搭配词统计数据

Collocates	Total occurrences	Co-occurrences with node	T-score
suicide	1445	118	10.847 712
crime	4405	46	6.708 760
murder	4418	29	5.292 233
offence	989	17	4.095 934
fraud	1322	12	3.420 872
conspiracy	633	11	3.295 005
adultery	198	9	2.992 524
robbery	524	9	2.980 215
sin	586	5	2.206 382
harm	1357	5	2.167 325

这些搭配词的一个重要特点是：它们在意义上是相互关联的，都是指生活中消极的或“坏”的事情；无论是“自杀”、“犯罪”、“谋杀”、“欺骗”、“抢劫”还是“伤害”都不是一般人期待或称道的事，都具有消极涵义；这个有限范畴的词汇不妨称为 *Crime* 类的词汇(crime category of words)。那么，*commit* 一词的重要搭配行为之一就是和这些意义关联的“消极涵义词汇”结合，表达一类意义。

1.2　节点词和几个意义不同的词搭配

在这种情况中，与节点词搭配的几个词项意义各不相同。例如，和 *criticism* 一词经常搭配的动词有：*make*、*receive*、*accept*、

attract、*deflect*、*dismiss*、*reject*、*lead to*、*be subject to*、*level at* 等，分别表示"进行批评"、"受到批评"、"接受批评"、"招致批评"、"转移批评"、"拒绝考虑批评"、"拒绝批评"、"引起了批评"、"常受到批评"、"对……进行批评"等不同意义。常和 *criticism* 搭配的主要动词的统计数据见下表。

表 6.2 *criticism* 一词的左动词搭配词统计数据

Collocates	Total Occurrences	Occurrences with node	T-score
levelled	299	28	5.273 584
rejected	2227	20	4.31 4221
faced	2858	20	4.269 477
attracted	1289	18	4.146 294
drew	1907	16	3.848 815
received	5998	16	3.524 485
accept	4595	15	3.496 749
drawn	2827	13	3.356 911
deflect	114	11	3.305 725
facing	2635	12	3.222 885
responded	1291	10	3.032 815
deal	12 065	15	2.885 113
dismissed	1528	9	2.838 482
renewed	839	8	2.734 361
subject	6754	11	2.670 848
led	7682	11	2.582 118
made	45 730	28	2.550 937

1.3 节点词仅和某个特定的词项搭配

在某些极端情况下，节点词的某个特定义项要求它仅与一个词搭配，表达特定的意义。*foot the bill* 这个词组就是这种情况。在该词组中，*foot* 的意义为"支付"(pay)，它要求 *bill* 为其独一无二

的搭配伙伴;*bill* 的同义词 *fee*、*charge* 等都不能替代它和 *foot* 构成搭配,表达"付账单"这一意义;本族语者是不说 * *foot the fee* 和 * *foot the charge* 的。在这种情况下,节点词的搭配力已至穷尽,其搭配范围也缩小到了极限。这种搭配在形式上已完全接近成语。之所以仍将其视为搭配,是因为从意义上来说 *foot the bill* 仍不是完全不透明的(opaque),而是一种半明晰(semi-transparency)的状态,也就是说,*bill* 一词的意义仍是明晰的。

柯鲁斯(Cruse 1989:41)将这种搭配称为粘着性搭配(bound collocation)。类似的搭配实例还有 *curry favour*、*highest explosives* 等等。在自由组合、有限组合(搭配)和成语组成的连续体上,粘着性搭配是介于成语和一般的词语搭配间的一种特殊组合,兼有两者的特点。

一般搭配是普通语言里词语组合的典型行为方式。从语言使用的创造性来说,词的任何组合或搭配都是可能的。然而,语言使用在很大程度上又是一种惯例式的行为(routine behaviour)。这种惯例式的行为反映在词语的使用上便是词语的典型搭配。典型性(typicality)是词语搭配的基本属性,它和真实语言使用中的自然性(naturalness)密切相关:没有典型性,就没有自然性;破坏了典型性,自然性也随之丧失。典型搭配是语言使用中的建筑砌块。本族语者在语言使用中主要的是根据语境和交际目的等因素选择那些典型的词语搭配来实现意义。他们选择 *commit offenses*,但是不会选择 * *commit defense*;他们选择 *commit conspiracy*,但是不会选择 * *commit planning*;同样的情形,他们 *make a criticism*,但不 * *do a criticism* 或 * *create a criticism*,他们 *accept a criticism* 或 *reject a criticism*,而不 * *recognize a criticism* 或 * *decline a criticism*。和成语一类的固定词组一样,搭配一类的半固定词组无疑是语言学习者需要掌握的最重要的内容之一,任何语言学习者不掌握大量的一般词语搭配,要想达到流利使用语言的水平是根本不可能的。词语搭配的另一个属性是其可描述性。这些典型搭配由于数目有限,是可以观察和描述的。而观察和描述这些搭配正是语料库语言学的重要内容之一。

2 修辞性搭配

2.1 修辞性义项:搭配标准与新颖性

修辞性搭配是基于词的修辞性义项组合而成的一种搭配。从根本上说,词语搭配是由于相关词的义项交互作用而结合在一起的。在修辞性搭配中,至少有一个搭配伙伴的义项是修辞性义项,其余搭配伙伴的义项是非修辞性义项。英语中的修辞性搭配俯拾皆是,如:

an ailing economy	a burning issue
a crushing blow	a scathing criticism
a barrage of criticism	a landslide victory
a sandwich course	a soap opera
the brain drain	kill the interest

修辞性搭配是语言使用者为追求新颖交际效果而创造性使用语言的产物。这种"创造性使用语言"的结果使得许多词都具有了一个或数个修辞性义项(figurative sense),由此也就有了修辞性搭配。然而,并不是所有的修辞性用法都是严格意义上的搭配。只有当一种修辞性用法在语料库中出现频率较高,也就是说,当它在语言交际中被反复使用时,才能算是一般所说的搭配。否则,只能算是一般的修辞格(figures of speech)的用法,如明喻(simile)、暗喻(metaphor)、借喻(metonymy)、提喻(synecdoche)和拟人(personification)等。一个修辞性用法是否能成为搭配,在很大程度上取决于相关词的修辞性义项是否被确立。当词的某个修辞性义项已被确立为正常语言使用的一部分时,它就会吸引一些具有一定特点的搭配伙伴(characteristic partner),并组成典型搭配。在这一

点上，统计手段可提供较为客观和实用的标准来确定词语组合是否为修辞性搭配。*run* 这个动词的搭配行为可用来说明该问题。该动词的一些修辞性义项已相当确立；这些修辞性义项的使用几乎和该词的一般义项 *go extremely fast* 一样普遍。因此，它的一些修辞性用法，如 *run a business*、*run an election* 和 *run a school* 等也就被普遍使用，以至成为典型搭配。表 6.3 中显示的是根据 COBUILD 语料库证据统计的有关信息。

表 6.3 *run* 一词的右名词搭配词统计数据

Collocates	Total occurrences	Co-occurrences with node	T-score
business	17 692	136	8.359 860
courses	3387	54	6.345 239
election	2031	61	5.850 798
club	17 839	92	5.543 490
company	20 432	99	5.480 202
services	9021	59	5.124 850
hotel	7788	54	5.041 661
campaign	7994	54	4.980 644
school	23 114	98	4.817 380
competition	6178	44	4.606 015

由表中信息可知，*run* 的名词搭配词大致可分为两类。第一类指称某种"组织"，如 *business*、*club*、*company*、*hotel* 和 *school*。和这类词搭配时，*run* 的意义是 *manage or control*，是一个修辞性义项。第二类搭配词指称某种"行动"或"重大活动"，如 *courses*、*election*、*services*、*campaign* 和 *competition*。和这些词搭配时，*run* 的意义是 *launch or start*，是另一个修辞性义项。这些修辞性意义似乎在对 *run* 的搭配范围起着制约作用；就目前讨论的情况而言，一个词要成为该动词合格的搭配伙伴，就必须属于上述所讨论的两个范畴，要么是"组织"范畴，要么是"行动"范畴。从另一方面来说，这些具体的搭配词也在对 *run* 的义项起着限定作用。这

也就是说，词语搭配的各个成分间存在着相互限定义项的关系。这种相互限定义项的关系影响着词的搭配力，影响着它的搭配范围。而只有当词的修辞性义项被确立，也就是被普遍使用时，其搭配力才可能较强，搭配范围才可能较大。

修辞性搭配的"新颖性"是另一个需要讨论的问题。一般来说，词的修辞性义项一旦被确立为正常语言使用的一部分，其新颖性也就大大减弱。因此，修辞性搭配的"新颖交际效果"也就远不如一般的修辞格，包括明喻、暗喻等等。但是，修辞性义项的确立是个程度问题，不可绝对化。有时，词的几个修辞性义项都已确立为正常语言使用的一部分，但"确立程度"不同。"不太确立"的义项所产生的词语组合会比"更为确立"的义项所产生的组合更富有"新颖性"，但不太会达到词语搭配的显著性标准。动词 *explode* 的搭配行为可说明这个问题。*explode* 的一般意义为 *blow up*。除此之外，它还有两个修辞性义项，一个是 *show something to be false or no longer true*，另一个是 *burst out violently*。COBUILD 语料库的证据表明，该词用于一般义项时的搭配词主要有：*bomb*、*battery*、*shell*、*fuses*、*mountain*、*nuclear device* 等。用于修辞性义项"show something to be false or no longer true"时的主要搭配词有：*myth*、*idea*、*thought*、*notion*、*theory* 和 *dissemination of copyright* 等表示抽象概念的词，基本上可概括为 *idea* 或 *thought* 这样一个语义范畴；用于修辞性义项 *burst out suddenly* 时的主要搭配词有：*rage*、*violence*、*tension* 等词。统计检验表明，第一个修辞性义项所产生的组合都不够显著，例如：

explode the {myth / thought / idea / notion / theory}

因此，它们最好被看做一般的比喻用法。但是第二个修辞性义项所

产生的组合都十分显著，例如：

rage
violence } exploded
tension

因此，它们最好被视为修辞性搭配。就“新颖性”而言，前一组的组合要比后一组稍强一点。但总的来说，无论是比喻还是修辞性搭配，都属于语言使用的“创造性”方面，都有一定的新颖性。这是修辞性搭配不同于一般搭配的一个特点。

词语组合的新颖性没有绝对标准，因为它与语言使用者的语言经验有一定的关系。由于语言经验的不同，一部分人认为“非常新颖”的词语组合，另一部分人可能觉得“毫无新意”。第二语言学习者认为“新颖”的组合，本族语者可能会觉得不以为然，甚至视之为陈词滥调。所以，在进行研究时，最主要的评判依据仍是统计测量的信息。除了 *explode* 一词的修辞性搭配，笔者还检查了 *raise*、*launch* 等词所组成的修辞性搭配。其中，只有为数不多的组合是显著的，现将有关数据总结在表 6.4 中：

表 6.4　有关修辞性搭配的统计数据

Node words	Collocates	T-score
raise	hope	6.449 862
raise	awareness	7.469 645
raise	consciousness	4.128 902
rage	exploded	2.434 772
violence	exploded	2.851 820
tension	exploded	2.018 723
launch	career	5.962 771
glittering	career	7.228 022
kill	interests	2.690 890
information	superhighway	8.511 835

2.2 修辞性搭配中的亡喻

亡喻(dead metaphor)是另一类重要的修辞性搭配。这是那些被反复使用的比喻。由于被语言社团长期使用,它们逐渐丧失了原有的新颖性,变成了亡喻。亡喻基本上已成为一种固定词组,形式上很像成语;其原有的新颖修辞效果基本上不复存在,很难再激活语言使用者的想象力。而语言学习者对付亡喻的办法也和对付成语差不多,即通过查阅词典来学习它们的确切意义和用法。英语中的亡喻很多,如 *a maiden voyage*(首航)、*a catch 22*(第二十二条军规)、*a hand-to-mouth economy*(勉强维持的经济)、*catch one's eye*(吸引某人的注意力)、*pay dividends*(产生回报)、*out of the woods*(脱离困境)、*abandon ship*(改换门庭)等。

亡喻在意义上是半明晰的,但组成亡喻的各词语已构成语义上的一个整体。因此,亡喻在形式上相当固定,其成分也不再有严格意义上的搭配范围。这是它与严格意义上的词语搭配重要的不同点所在。在下列各组词语组合中,a 类是亡喻,b 类则是一般搭配:

1a abandon ship

1b abandon (a) {hope / plan / project / one's wife}

2a carry weight (to be influential, important)

2b carry (a/the) {burden / weapon / message / news / information / arms}

3a bear the brunt of (endure the worst part or main strain of)

3b bear(a/the) {burden / responsibility / punishment / cold(heat) / resemblance}

4a draw the line (to fix a limit)

4b draw a/an {conclusion / inference / distinction / breath / analogy / lesson}

当然，上述分析和区分不是绝对的，只是个程度问题。对语料库研究者来说，重要的是在定量分析的基础上，对词语组合进行适当的定性分析和描述，以利于语言使用和语言学习。

3 专业性搭配

专业性搭配是专业技术与学术研究领域里出现频率很高的词语组合，是专业文本里承载思想与概念和组织语篇信息的重要词语单位。这些词语组合在普通语言文本里使用得较少，但它们是科技和学术人员表达具体的专业概念、思想和活动，赖以交流的重要词语手段。专业性搭配这一概念和韩礼德(Halliday)的语域思想是一脉相承的。韩礼德将语域分为域(field)、旨(tenor)和式(mode)。根据他的思想，语域的三个方面因素，域、旨和式，对语言形式的选择具有决定性的影响；三者中任何一个发生变化都可能导致交际者选用不同的句法结构和词语形式(lexical pattern)等。由

于学术研究和科技领域特有的域、旨和式，一些特有的词语组合常被频繁地使用，并形成典型的意义和形式特征。研究这些形式和意义，便是词语搭配研究者的重要任务之一。

就其意义而言，专业性搭配大体上实现两大类意义。其一是概念意义，表达有关研究领域里的重要概念、思想和方法，或者表述某种实践、经历或活动。其二是篇章意义，主要是起着篇章信息的组织作用，或陈述研究者的观点、证据和研究结果。比如，*perform an experiment* 这个词语搭配实现的就是概念意义；它表述有关研究领域里的实践活动。而 *experiment indicates* 这个搭配则主要实现语篇意义，陈述研究者的观点和证据(详见第三章中的2.2.4一节)。专业性搭配还有其典型的形式特点、结构转换机制和内部意义关系。这些将在下面予以讨论。

3.1 专业性搭配的形式特点

专业性搭配有许多典型的构成方式。首先，在专业性搭配中，名词化的使用形式(nominalization)极为普遍。例如在 N + N 搭配中，名词化的形式出现频率就很高。在有些情况下，第一个名词是名词化形式，如 *application needs*、*application problem*。在有些情况下，第二个名词是名词化形式，笔者以“data”为节点词，在JDEST语料库中进行了查询，结果发现与节点词搭配的名词大多数是名词化形式，如 *data collection* (JDEST数据)

1	p optimization for comfort control; plus	data collection of kW and kWh for monitoring
2	o further improve the efficiency of the	data collection by computerizing this evalua
3	be inadequate since all of a session's	data collection would have to be discarded if
4	everal steps are involved. These include	data collection and preparation, matching of
5	he source of the problem based on later	data collection. In order to have team member
6	been controversy as to the accuracy of	data collection on machine tools; however,
7	ssary to converge on the nominal shape.	Data Collection On-Machine Data acquisition
8	posite to an online operation. The local	data collection point — a sales office, a fa

9	orm the weld without human intervention.	Data collection is done by the 3-D vision sys
10	t of time that must be invested before	data collection can begin), they can collect
11	d thereby improve the efficiency of the	data collection. A third objective of this st
12	ment of information as better methods of	data collection, interpretation and more arti
13	uses of that delay. This is part of the	data collection process and will be consta
14	er the entire Gulf, and System Argos, a	data collection system to track the velocity
15	rogress of product improvement. The CIM	data collection systems which are used in AT&

除了 *data collection* 之外，与 *data* 这个节点词有关的搭配，还有 *data acquisition*、*data analysis* 和 *data processing* 等。

名词化形式使用极多，这一现象还反映在"nominalized item + of + N"这种搭配序列中，在由学术论文(research articles)组成的一个子语料库中(sub-JDEST)，这种搭配方式极为普遍。比如，"investigation"一词总共在该库中出现了 129 次，其中有 36 次是出现在这种序列中，占 27.99%。由于篇幅有限，下面只显示其中的 15 行词语索引：

1	ABSTRACT Based on a field	investigation of 500 engineering managers, specifi
2	can provide fertile ground for the	investigation of individual predictors of innovati
3	he end result of extensive inhouse	investigation of alternatives to the purchase of
4	the indirect fuel ell systems. The	investigation of tri-fluoromethanesulfonic acid
5	ight profitably be utilized in the	investigation of short-cycle operations that are
6	Such an approach requires a complex	investigation of technological process and, first
7	may form (30-34). Jones (31), in an	investigation of the hot cracking susceptibility i
8	st. Often such a choice entails an	investigation of equipment costs, with the relativ
9	suggesting that a more direct	investigation of this would be relevant today. Som
10	discovered it in the course of an	investigation of the properties of water waves
11	city could be determined. The	investigation of the different re-combinations of
12	these characteristics support the	investigation of the lognormal distribution, a third
13	cited mechanical system and the	investigation of the stability of structure belongs
14	following a perturbation, an	investigation of stability by means of dynamic analy
15	to be analysed sub In an elastic	investigation of the overall problem, it is always

从上述词语索引可以看出，该搭配序列中的关键词 *investigation* 之前可用不定冠词*an*，如 *an investigation of costs*；也可用定冠词 *the*，如 *the investigation of manufacturing system*，或者其扩展形式 *a complex investigation of technological process*，还可用零冠词，如 *investigation of the stability* 等。一个搭配序列的出现频数占了其节点词总频数的 20%以上，这一事实足以说明其重要性。对语料库研究者来说，需要进一步研究的是该搭配序列中介词后面跟的具体名词，以便搞清楚该序列更加典型性的内容，如 *problem*、*process*、*costs*、*alternatives*、*properties*、*stability* 等。除了"investigation + of + N"这一搭配序列外，还有"understanding + of + N"、"examination + of + N"、"application + of + N"、"discussion + of + N"、"expression + of + N"、"analysis + of + N"、"study + of + N"和"survey + of + N"等。它们在 JDEST 语料库中出现的频数分别占关键词总频数的 57.4%、48.7%、41.9%、41.8%、24.1%、23.9%、23.2%和 21.9%。

专业性搭配另外一个引人注目的特点是 V-ing 形容词和 V-ed 形容词在 ADJ. + N 搭配中的大量使用。由 V-ing 形容词或 V-ed 形容词加上名词构成的搭配在普通英语中也时有出现，但远不如学术英语中使用的普遍。一些 V-ing 和 V-ed 形容词在学术英语中经常与有关名词结合构成名词性词组、反复地使用，而这样的用法在普通英语中却并不十分突出。例如：*increasing* 和 *increased* 两个形容词的搭配情况就是如此。在 JDEST 语料库中，*increasing amounts* 和*increased amounts* 这样的搭配极多：

(1) ***Increasing amount /s***

1	decreasing green density signifies an	increasing amount of internal surface area
2	1, strengths increase gradually with	increasing amount of deformation at high deforma

3	rocket, NASDA gradually introduced an	increasing amount of Japanese technology and aut
4	physics texts because of the ever	increasing amount of material which modern phys
5	increase in the green density with	increasing amounts of lubricant. Some data on th
6	g well, the produced gas will contain	increasing amounts of the non-combustible flue g
7	uring and after deformation of. With	increasing amounts of deformation , dislocation
8	at recovery is easier to proceed with	increasing amounts of deformation. Coldren, Eldi
9	one and the same amount of water but	increasing amounts of sand. However if there is
10	was shown in the following way. When	increasing amounts of enzyme are preincubated fo

(2) *Increased amount /s*

1	is used for both K and K2. A greatly	increased amount of data on column behaviour has
2	el has a reduced hardness owing to an	increased amount and coarser grain of retained a
3	he urine. During acute pancreatitis,	increased amounts of amylase are excreted into t
4	e angiotensin leads to production of	increased amounts of aldosterone. Therefore sodiu
5	Propeller Skew The concept of using	increased amounts of blade skew in an attemp
6	a set of propeller blades having	increased amounts of skew. Consequently, a new
7	lymphocytes are stimulated to produce	increased amounts of immunoglobulins, and fibro
8	cy. Researchers have found that	increased amounts of a second pancreatic hormon
9	method Generally the adoption of	increased amounts of skew would seem to offer
10	as up a factor of 50, consistent with	increased amounts of ClO in the presence of brom

类似的搭配还有:

increasing {
- cost
- demand
- complexity
- importance
- interest
- pressure
- use

increased {
- cost
- complexity
- demand
- efficiency
- levels
- risks

这些搭配在 JDEST 语料库中出现的频数很高,但在 LOB 语料库和 BROWN 语料库中的频数却很不显著。这两个语料库也有 *increasing* 和 *increased* 与一些名词结合的实例,如 *increased use* (*expenditure*、*wages*、*speed*、*profits*、*investment*、*expenses*、*charges*)。但 *increased use* 和 *increased expenditure* 各出现了 2 次;其他组合各出现了一次。从定性分析的角度看,其中的一些组合可以视为搭配。但是,对多数组合而言,要确定它们是否搭配,定量分析是必不可少的重要手段。而在 JDEST 语料库中,V-ing + N 或 V-ed + N 型的组合比比皆是,如 *supporting evidence*、*supporting argument*、*growing demand*、*working environment*、*accepted value*、*accepted practice*、*fixed distance*、*fixed rate*、*improved methods*、*improved performance*、*proposed model*、*required performance*、*established principles* 等等。这些组合的频繁使用无疑表明它们是学术研究人员表达意义的重要词语手段,而这种搭配形式也无疑是专业性搭配的重要特点之一。

普通英语中每一类搭配在学术英语中都会有其具体的实现特点。有时,实现形式可能是具体的个别现象,不一定具有普遍意义。但它们依然应加以描述。如,PREP + N 是普通英语中常见的一个搭配类别。当然在学术英语中也很多。然而,学术英语中的这类搭配却有其自身的实现特点。比如,在普通英语中,如果要表达"方式"、"手段"、"方法"等意义,人们常用 *by means of* 这个词组。在学术英语中,*by means of* 用得也很多。然而学术论文常有专门描述"研究方法"的一节,称为 *method section*。在 *method section* 一节中,描述一般的、概括性的方法与具体的方法、描述本领域里较为知名的方法、研究者在论文中业已提及过的方法等,用的词汇手段都有区别。其中,用得较多的是 *by use of* 这一词语搭配。*by use of* 在 JDEST 语料库中共出现了 51 次,下面是其中的 15 行词语索引:

1 he helical spring); and "fine control" by use of the piezoelectric crystal tube, wh

2	ncs/cm during the course of an experiment	by use of the clamp-adjusting rod. The dista
3	ed a number of approaches to recover oil	by use of surfactants to reduce interfacial
4	m has been designed to upgrade basic data	by use of statistics. curve match, and simil
5	; optimal problems are most easily solved	by use of maximum principle. When a rigorous
6	and mud filtrate must first be determined	by use of Figures 6 or 7 at the appropriate
7	unt importance and is usually attainable	by use of moist heat or ice compresses; rest
8	ant (Ke) was determined for each subject	by use of a model-dependent graph strippin
9	ch analyed for tobra- mycin concentration	by use of an immunofluores- cence assay proc
10	ra- mycin(7.5mg/kg/dose) was administered	by use of a nebulizer (Microstat Ultra Nebu
11	on the purified PCR product (-0. 2 pmol)	by use of a commercial kit (Se- quenase vers
12	imization is performed in multiple stages	by use of orthogonal projection. At each sta
13	previously estimated risk error of 1.5%.	By use of this method, the 25 at-risk subjec
14	5%) to maximise the take-off performance	by use of the flap effect of the elevons. Th
15	onal "white lies." identification, or	by use of rationalizations. Normally suc

从这些词语索引可以看出，*by use of* 常被用来详细描述较为具体的方法，*use* 一词右面的搭配词一般指具体的工具、事物、器械、图表等，如 *the piezoelectric crystal tube*（1）、*the clamp-adjusting rod*（2）、*surfactants*（3）、*Figures 6 or 7*（6），*moist heat*（7）、*model dependent graph*（8）、*orthogonal projection*（12）等等。该词组也可用来介绍一般的和笼统的方法；*use* 右边的搭配词也可以是表达抽象概念的词，如 *statistics*（4）、*maximum principle*（5）、*this method*（13）、*rationalization*（15）等。

by use of 的这些用法基本上是学术英语特有的，在普通英语文本中出现的实例极少。在 LOB 语料库，该词组仅出现了两次。即使这两个实例也是出现在科技文章中，而不是普通英语文章中。下面是 LOB 语料库中这两个实例的词语索引：

1	'Cutting can be carried J70 182 out	by use of a thin diamond slitting wheel or by us
2	e of a thin diamond slitting wheel or	by use of an ultrasonic J70 183 machine with a k

在上面的两行词语索引中，*by use of* 右面的搭配词 *diamond slitting wheel* (1)和 *ultrasonic machine* (2)表明它们所出现的文本是学术英语文本而非普通英语文本——LOB 语料库基本上是个普通英语语料库，但也包含了少量的学术英语文本。

专业性搭配的形式特点还很多，不可能在这篇短文中面面俱到地加以讨论。对研究者来说，对这些具体的专业性搭配的形式和意义做出准确的描述和概括并非易事。然而，这正好说明了进一步加深研究专业性搭配的必要性。

3.2 专业性搭配的结构转换

词语搭配行为要限定在具体的语法框架或类联接内研究。类联接代表了搭配类别。然而，搭配类别之间是可以相互转换的。就专业性搭配而言，其搭配类别之间的转换呈现着一定的规律和特点。这里面有句法制约的作用，也有语境和意义表达的制约作用。

3.2.1 N + V 搭配与 N + N 搭配的转换

在学术英语中，N + V 搭配与 N + N 搭配的转换较为普遍。我们以词形 *performed* 的搭配行为为例进行分析。在 JEDST 语料库中，与 *performed* 搭配，作为其主语的名词主要有 *engines*、*machines*、*computers*、*tools*、*systems*、*procedures*、*software* 和 *unit* 等词，根据子语料库 sub-JDEST 库中的数据统计的有关结果如下：

表 6.5 *performed* 一词的左名词搭配词统计数据

Collocates	Co-occurrences with node	Total occurrences	Z-score
engines	3	54	9.75
machines	9	476	9.27
computers	14	1190	8.59

（续表）

Collocates	Co-occurrences with node	Total occurrences	Z-score
tools	6	321	7.51
systems	13	1614	6.34
procedures	4	262	5.43
softwares	4	266	5.37
unit	4	602	3.02

然而，上述 N ＋ V 搭配一般都可转换为相应的 N ＋N 搭配。转换的一般规律为：原 N ＋ V 搭配中的 N 多数要变为单数形式，在 N ＋ N 搭配中充当第一个 N，原 N ＋ V 搭配中的 V 要变为相应的名词形式，在 N ＋ N 搭配中充当第二个 N。下面是 JDEST 语料库中有关的几个 N ＋ N 搭配（这些都是显著搭配，限于篇幅，具体数据不再列出）：

engine
machine
computer
system
} performance

当然，并不是所有的 N ＋ V 搭配都可以转换为相应的 N ＋ N 搭配。表 6.5 中的 *tool* ... *performed*、*procedures* ... *performed*、*softwares* ... *performed* 三个显著搭配就没有相应的 *tool* ... *performance* 等 N ＋ N 搭配。但是证据表明 *student* ... *performance*、*worker* ... *performance*、*employee* ... *performance* 和 *company* ... *performance* 是显著搭配，Z 分值分别达到了 6.52、8.61、6.58 和 2.42。但在相应的 N＋V 搭配中，*students* ... *perform* 等却不是显著搭配。（这并不是说 *students* ... *perform* 不是搭配，仅仅是就 JDEST 语料库的证据而言，该组合不够

显著，在其他语料库里，它可能是个显著搭配。）这说明，任何规律都有例外情况，对词语搭配而言，尤其如此，因为词语搭配研究的毕竟是词语的个体行为。

3.2.2　V+N 搭配与 N+N 搭配的转换

学术英语中的 V + N 搭配与 N + N 搭配也可发生转换。一般规律为：V + N 搭配中的 V 要变为相应的名词形式，充当 N + N 搭配中的第二个 N；V + N 搭配中的 N 变为相应的单数形式或者不变，充当 N + N 搭配中的第一个 N。仍以 *performed* 这一词形的搭配行为来说明。JDEST 语料库中有许多 *performed ... experiment*、*performed ... operations* 这样的显著搭配，详见下表数据：

表 6.6　*performed* 一词的右名词搭配词统计数据

Collocates	Co-occurrences with node	Total occurrences	Z-score
experiment	17	174	31.19
operations	20	494	21.25
tests	15	465	16.25
calculation	6	97	14.44
task	6	231	9.10
analysis	4	143	7.81
elements	8	692	6.42
functions	5	368	5.64
analysis	12	964	4.71
work	11	1778	4.71
activities	4	354	4.47
operation	6	810	4.03
function	4	712	2.61
research	4	737	2.52
test	3	872	1.91

转换后相应的 N + N 搭配为：

task
operation
analysis
research
test
work
} performance

同样，并不是所有的 V + N 搭配都有相应的 N + N 搭配形式。如，表 6.6 中的"performed ... function(s)"、"performed ... calculation"和"performed ... activity"就没有相应的"function performance"、"calculation performance"和"activity performance"。这就是说，研究者可以参照搭配类别转换的一般规律去观察词语行为，但更重要的是以语料库证据为准对词语行为进行描述。脱离证据的"概括"和"抽象"是与语料库语言学的原则格格不入的。

3.2.3 N + of + N 搭配

在很大程度上，上述三个搭配类别都可变为 N + of + N 搭配。具体地说，*computer ... performed* (N + V)可变为 *computer performance* (N + N)，也可变为 *the performance of computers* (N + of + N)，*performed ... task* (V + N)可变为 *task ... performance* (N + N)，但也可变为 *the performance of the task*。在 JDEST 语料库中，"performance + of + N"的搭配实例极多。现引举若干个有关的例子如下：

the performance of systems
the performance of computers
the performance of the job

the performance of the test
the performance of the student
the performance of the engine
the performance of the operation
the performance of tasks
the performance of the machine
the performance of the work

3.3 专业性搭配的内部意义关系

专业性搭配有其一套内部的逻辑语义关系，大致上包括：施动者—动作、动作—施动者、动作—受动者、受动者—动作等关系。

施动者—动作(agent-action)关系。施动者(agent)即动作的发出者，如 *machine* ... *performed* 中的 *machine* 即是施动者，而 *performed* 则表示具体的动作。具有这种逻辑语义关系的专业性搭配有：*computers* ... *performed*、*systems* ... *performed*、*system performance*、*student performance* 等等。

动作—施动者(action-agent)关系。上类关系的两个要素刚好在这类关系中作了位置上的调换，这是搭配在结构上转换的结果。下列搭配都具有动作—施动者的关系：*the performance of the machines*、*the performance of systems*、*the performance of the worker* 等等。

动作—受动者(action-patient)关系。动作可由具体的动词或名词表示；受动者(patient)即动作所作用于的事物，如 *performed* ... *tasks* 中的 *performed* 一词即表示具体的动作，而"task"则是动作的受动者。下列搭配都具有动作－受动者关系：*the performed tests*、*the performed tasks*、*the performance of the test*、*the performance of the operation*

受动者—动作(patient-action)关系。上类关系中的两个要素在这类关系中作了位置上的调换，这也是搭配类转换的结果。下列

搭配具有这种逻辑语义关系：*task performance*、*operation performance*、*research performance*、*test performance*。

上述仅仅是专业性搭配内部所具有的部分逻辑语义关系。另外还有其他多种多样的逻辑语义关系。如 *required performance* 可以被认为体现了一种目标—动作(goal-action)关系；该搭配可以解释为 *perform in such a way as to reach the required standard*。另外，许多搭配明显地表达了某些重要的专业概念和思想，如 *acceptable performance*、*potential performance* 和 *overall performance*。还有一些专业性搭配明显揭示了语域因素，如 *technical performance*、*financial performance* 等。这些搭配最好从它们所实现的具体概念意义和具体语篇意义上去分析、描述。这一节对专业性搭配的语义关系所作的分析只是一点尝试性的工作，旨在探索一种可能的研究方法。

4 惯例化搭配

惯例化搭配是一种具有高度因循性的词语搭配，被语言社团(speech community)或者话语社团(discourse community)视为某个特定意义的标准表达形式而沿用。弗斯(1957:187)曾经提醒人们注意研究那些惯例化的句子组成部分。他区别了一般的搭配和扩展搭配(extended collocation)。他说的扩展搭配大致就是本节要讨论的惯例化搭配。在有关词语研究的文献中，人们常用各种各样的说法来指这类搭配，如形式—意义配对(form-meaning pairings)(Pawley and Syder 1983)、词语片语(lexical phrases)(Nattinger and DeCarrico 1992)、半预制性词组(semi-preconstructed phrases)(Sinclair 1991)、公式块(formulaic chunks)(Widdowson 1989)。这说明惯例化搭配这一语言使用现象早已引起语言学家们的注意，值得进行研究。

惯例化搭配有其形式和意义上的显著特点，使之有别于一般的词语搭配。首先，惯例化搭配的长度不等、结构层次多样。它可以是词组级(phrase level)的一个搭配序列，也可以是分句级(clause level)的搭配序列，还可以是句子级(sentence level)的搭配序列。可以只由两个词构成，也可以由六七个词构成。组成惯例化搭配的词及其语法形式也基本上是固定或半固定的，即使有若干个变体也绝对没有一般搭配中的"搭配范围"所允许的自由度。它所表达的意义也是一种为语言使用者都认可的"标准意义"，不可能有歧义等。

惯例化搭配有两种。一种是被语言社团成员接受并沿用、表达某个特定意义的词语搭配序列，另一类是被某个话语社团反复使用并表达特定意义的搭配序列。

4.1 普通语言里的惯例化搭配

所谓惯例化(institutionalization)又可叫词语化(lexicalization)。波利和塞德(Pawley and Syder)曾从意义的权威性和标准性、形式的因循性和结构的任意性三个方面来界定惯例化。他们认为，如果一个句子或句干符合下述三个条件就词语化了：其一，它表达一个文化上授权的意义或标准的概念；其二，它被认可为表达该意义的标准形式；其三，在语言学上，选择它作为一个标准表达方式是一种任意性行为。("a sentence or sentence stem is lexicalized if it (i) denotes a meaning which is culturally authorized, i.e. is a standard concept in the speech community; (ii) is recognized to be a standard expression for the meaning in question; (iii) is an arbitrary choice (or somewhat arbitrary choice), in terms of linguistic structure, for the role of standard expression")(1983:211)。

据此，一般语言里的惯例化搭配的界定标准应当是：1)它是表达语言社团里某个固定意思的词语组合序列，其意义清晰明白，不

模棱两可；2）这种词语组合序列被语言社团成员认可为表达该固定意思的固定或半固定形式；3）从语言学角度来看，选取该词语组合序列来表达该固定意义是一种任意性行为，也就是说，与其他同义或近义的表达方式相比，该序列本身并没有特别的因素使其选为近乎标准化的表达方式。如，英语中表示“死罪”这一意思的词语是 *capital offense*。*capital offense* 表示的就是必须要判处死刑的那种严重罪行。这一概念清晰明了，没有歧义；本族语者将其认可为表达“死罪”这一概念的固定或半固定形式，也就是说，人们在表达“死罪”这一概念时，大都沿用 *capital offense* 这一说法；因语境不同，也会用 *capital crime* 这一说法，所以应称之为半固定的表达方式。从语言学理论上来讲，*capital offense* 这一词组并没有什么比 * *death offense*、* *death penalty crime* 或 * *death sentence offense* 等说法特别的地方，使其选为这一概念的标准表达形式。无论从意义上还是从形式上说，惯例化搭配都是一种具有高度因循性的词语组合序列，是地道语言使用的重要内容。

惯例化搭配既可表示概念意义，又可表示语篇意义。普通英语里有大量的固定词组和半固定词组，实质上都属于惯例化搭配的范畴。它们或长或短，可以是词组级的、分句级的或句子级的。

(1) 词组级的：

a close vote（几乎势均力敌的选举，投票险胜）
a forced landing（紧急迫降）
a hot line（热线电话）
a vicious circle（恶性循环）
ladies and gentlemen（女士们、先生们）
the lay of land（地势）
by pure coincidence（纯属巧合）
in broad day light（光天化日之下）
point taken（接受你的观点）
under the circumstances（在这种情况下）

(2) 分句级的:

To cut a long story short ... (长话短说……)
To go ahead(开始)
If you see what I mean (如果你明白我的意思)
Having said that ... (说了这些之后,那么……)
To see how the land lies (察看地势)
Let's see if we can ... (我们看一下能否……)
To put it another way ... (换个方式说……)
To make matters worse ... (更糟的是……)
There is the possibility ... (有可能……)

(3) 句子级的:

That's the point. (这就是要点所在。)
That's only a point of time. (这只是个时间问题。)
I haven't the foggiest idea. (我一点也不知道。)
Business is bad. (生意不好。)
What is going on here? (你们在干什么?)
Let's call it a day. (今天到此结束。)

4.2 学术英语中的惯例化搭配序列

要界定学术英语领域里的惯例化搭配,必须参照"话语社团"这一重要的概念。话语社团不同于语言社团,它和人们所从事的专业活动密切相关。斯韦尔斯(Swales)(1990)讨论了话语社团的重要界定特点,是迄今为止关于"话语社团"这一概念最为详尽的论述。按照他的界定体系,在各自领域从事研究工作的科学家和学术研究人员都在不同程度上隶属于一定的话语社团。一个话语社团有大体上一致的和公开的共同目标,比如计算机科学家和语言学家有各自具体的研究目标;有供社团成员交流信息的机制,比如专业研究会、协会、学会等组织,和研讨会、信息通报、通讯往来等活动;有一两个为促进目标实现而用于思想交流的文类形式或语篇发展形式,

以制约人们选取适当的专题、采用得体的形式、完成一定的功能以及决定语篇要素的位置等，比如各学科领域办的期刊、学报等所规定的论文的形式与结构；它还有一套专用的词语系统，比如技术词、次技术词及其特有的用法。因此，话语社团形成了一套高度因循性的交流方式，包括文类形式、语篇模式和词语的选择与组合。这些方式是话语社团特有的，外行人员通常对之不甚了解。

按照这些重要的学术思想和观点来考虑，学术英语领域里的惯例化搭配不妨参照下列标准界定：

(1) 它是专业研究领域里高度因循性的词语组合序列，用于表达特定的概念或实现一定的意义；

(2) 它为话语社团成员，也就是专业领域里的同仁，认可为表达某个特定意义的固定或半固定的词语组合形式；

(3) 从语言学理论上讲，选择某个词语组合形式来表达某个特定的意义，或多或少都带有某种任意性；也就是说，被选择的形式本身并没有什么特别之处使其选为大体上固定的表达方式。而且，外行人员一般都不熟悉这样的高度因循性的词语使用形式。

学术英语领域里的惯例化搭配与普通英语中的惯例化搭配的不同之处在于它表达的是一个专业概念或专业意义；熟悉这种惯例化表达方式的人员是一个有限的群体，基本上是专业同行。从这个意义上讲，研究这一类惯例化搭配似乎显得更为必要和重要，因为说到底，语言学研究人员与自然科学领域里的研究人员不属于同一话语社团，因此也就不了解这一类惯例化词语手段。那么，我们有理由认为，尚有大量的惯例化词语搭配序列虽被科学家沿用已久，但语言学家对此却并不十分熟悉和了解，或者说，这些被科学家沿用已久的惯例化词语搭配序列仍未被语言学家重视、发现、研究和描述——而这正是词语搭配研究的重要和艰巨的任务之一。

同普通的惯例化搭配一样，学术英语中的惯例化搭配也是长度不等，结构层次多样。现在，我们讨论几个实例。

第一个实例是“(a) case in point”。该搭配意为“恰当的例子”，常被研究人员用来解释和说明正在讨论的问题。如：

The mechanics of plastic fracture are inherently different from those of brittle fracture — instability effects are a case in point. (JDEST 数据)

case in point 可被用于各个专业领域，政治、经济、教育、机械、新闻等等，这可由下面的索引看出：

1	diplomatic moves. The most notable case	in point concerns the events leading to the
2	cture — instability effects are a case	in point. Ductile fractures follow from an i
3	ng English perception - verbs as my case	in point. If we are willing to look at such
4	he preparation of lunar tables is a case	in point. In 1937, Howard Aiken. a gradu
5	insufficient and misleading data. A case	in point is the comparison of uniform mesh
6	ter the evaluation. A most dramatic case	in point is Rader's 1968 paper, "The DFT whe
7	umerical solutions are realizable. A case	in point is the distillation column. Typical
8	filtering on every measurement. A case	in point is the Natural Gamma Ray Spectromet
9	anges in negative self-feelings. A case	in point is a study in which subjects who
10	ade no sense. They suggested, as a case	in point, that if they could do the budget
11	n or convocation ceremony is a good case	in point. The standard of education and pers
12	etitiveness of events. Pecheux is a case	in point though he represents an advance on
13	ment as a political journalist is a case	in point. ~ T2520C91B1979SL Intercommunicat

这个词组可作为主语使用，如 *A case in point is* ... (9)。又可作表语，如 *Pecheux is a case in point*(12)，还可以作插入成分，如 *as a case in point* (10)。一般来说，无论作何种成分，其典型的用法是 *case* 之前不用形容词性修饰语，形式较为简炼，如上述大多数索引所显示的那样。这符合学术英语的特点。只有在极少数情况下，*case* 之前加用形容词性修饰语，如索引行(1)、(6)和(11)所用的 *the most notable case*、*a most dramatic case* 和 *a good case* 等情况。

第二个例子是 *There is ... evidence to suggest*。该搭配序列意为"有证据表明"。在 JDEST 语料库中，*evidence* 一词总共出现

了 891 次。在 891 个实例中，*evidence* 与其他各种各样的词组成搭配，比如 *evidence shows* 等 N + V 搭配、*statistical evidence* 等 ADJ + N 搭配、*field evidence* 等 N + N 搭配。另外还常用于存在句 *There is evidence* 结构中，这个“存在句”结构共有 88 例，如 *There is evidence for* …、*There is evidence of* …、*There is evidence that* …和 *There is evidence to* …。*There is evidence to* …这一结构形式共有 36 例，其中不定式中的动词有 *confirm*、*suggest*、*conclude*、*support* 等。下面是其中的 18 行索引：

1	tt welds are concerned there is little	evidence to confirm the latter statement, but s
2	person's? There is no physiological	evidence to suggest that the senses of touch, t
3	reas. However, there was insufficient	evidence to conclude that RFLP types were rest
4	drian (1981) suggests that there is no	evidence to show that biological aging causes d
5	vice environments, there is	evidence to suggest that pressure testi
6	concluded that there is not sufficient	evidence to suggest that the pebbles were deriv
7	ifferent formative processes. There is	evidence to suggest that the extension in a num
8	by the formula Fe-3C, there is little	evidence to suggest that a true chemical com
9	misleading, for in some cases there is	evidence to show that acid bodies were formed b
10	Further, there is	evidence to indicate that the pile driving ca
11	et even in Wolverhampton there is some	evidence to suggest that the parish constable w
12	e is not quite so clear. There is some	evidence to suggest that difference is preferab
13	fully adequate, there appears to be no	evidence to support the view that, once the ext
14	ated. In contrast, there is sufficient	evidence to suggest that externalization is wid
15	tubers is determined. There is no good	evidence to suggest that under normal condition
16	ure 6.5). However, there is no strong	evidence to suggest they form complexes. Quite
17	ng work. In summary, there is ample	evidence to suggest that the microalgae compr
18	t are connected. However, there is no	evidence to link the two processes. Moreover, f

由上述词语索引可知，*evidence* 之前有不少形容词修饰语，*enough*、*ample*、*good*、*sufficient*、*little* 等。这些形容词修饰语或多或少都表达了研究者的某种态度。在索引行 3、10、14、16、17 和

18 中，该搭配序列紧紧跟着 *however*、*further*、*in contrast*、*in summary* 这样的语篇连接手段。动词 *suggest*、*indicate*、*support* 等之后的词语表明该搭配序列是被用来陈述研究者的某种观点，这一点可由下列更为详细的引句证实：

1. There is evidence to suggest that the extension in a number of western Pacific marginal basins is also different from normal spreading.（JDEST 数据）

在引句 1 中，*There is evidence to suggest* 显然是被用来陈述作者的观点：西太平洋一些边缘盆地的地质形成过程中的延伸不同于一般的地质延伸。当作者反驳或不同意某种观点时，则常用 *There is no*（*little*）*evidence to suggest* ... 等表达不同观点。下例中的 *There is no strong evidence to suggest* 与 *However*、*Quite to the contrary* 等语篇连接语紧密交织在一起，陈述作者相反的观点。

2. Unlike the respiratory system, the major membrane complexes in the photosynthetic system are present in approximately equimolar amounts. However, there is no strong evidence to suggest they form complexes. Quite to the contrary, this system displays dramatic lateral heterogeneity that results in photosystem II being localized ...（JDEST 数据）

事实上，*evidence* 的形容词修饰语 *enough*、*good*、*sufficient*、*ample*、*strong*、*some*、*insufficient*、*little*、*no* 等都是在不同程度上强调所陈述的观点，或者说，增强"陈述观点"的力度。

所以，*There is ... evidence to suggest* 这个搭配序列是研究者就某些具体数据、事实和研究情况表明自己学术观点或态度的一

个词语组合。他可以用不同的形容词来修饰 *evidence* 一词，从而增强自己陈述观点的力度。从 18 行词语索引的情况来看，这个搭配序列中还可用 *suggest* 之外的一些动词，如 *confirm*、*indicate*、*show*、*conclude*、*support* 等。但 *suggest* 是使用频率最高的动词，占 61%(11/18)。在整个语料库中，由于 *evidence* 是个出现频数很高的词，所以这一搭配序列所占的比例并不高。但在 88 个“存在句”实例中，该序列占了 40.9%(36/88)，比例相当高。这就说明，这是一个具有高度因循性的惯例化搭配，学术研究人员们都视其为“呈述观点”的近乎标准的方式之一，大家都沿用它来表达学术观点。外行人员对此并不十分熟悉，可能会觉得该序列无非是个存在句，和英语中不计其数的存在句一样，没有什么特别之处值得研究和描述。然而，上述分析无疑表明，惯例化搭配不仅在学术英语中表达具体的专业概念和思想，而且起着组织语篇信息的重要作用。这些都是语言学研究不可忽视的重要内容。

4.3 惯例化搭配序列在学术论文中的语篇功能

惯例化搭配序列在学术论文中发挥着重要的语篇功能。我们以学术论文的引语部分为例来讨论这个问题。引语的三个语步(moves)(见 Swales 1990: 141)：确立区域(Establishing a territory)、确立位置(Establishing a niche)和占领位置(Occupying the niche)分别都有各自的惯例化搭配序列来组织语篇信息、揭示语步间的衔接。

4.3.1 *Of particular interest*：声称研究课题在研究领域中的中心地位

在“确立区域”这一语步中，研究者一般都要首先声称自己所研究的课题属于一个重要的领域，是处于中心地位的研究内容，具有很高的价值等等，以吸引话语社团对论文中所报道的研究项目的注意。斯韦尔斯认为，“声称中心地位是在吁请话语社团成员

注意研究论文中所报道的研究内容，并将其接受为充满活力的、有意义的或业已确立的研究领域的一部分”（“Centrality claims are appeals to the discourse community whereby members are asked to accept that the research to be reported is part of a lively, significant or well-established research area.”）(1990: 144)。

在“声称中心地位”(claiming centrality)这一“阶”中，*of particular interest* 就是重要的语篇组织手段：

1	separate organisations. It is	of particular interest and relevance to note that the in
2	p engines are discussed.	Of particular interest are the results obtained from en
3	or fixed structures.	Of particular interest are the wave loads on fixed stru
4	.21 and 7.22. The latter is	of particular interest as the different roles of nuclear
5	allowed to vary. This case is	of particular interest because in his commentary on De R
6	r to research further points	of particular interest. The book will serve primarily as
7	f property. This is a fact	of particular interest, for Macfarlane has argued that
8	and mountain ranges, and are	of particular interest in the study of the cloud-shrou
9	ican poetry. Riddel's case is	of particular interest in that he began his critical car
10	stance and a group of alloys	of particular interest in the present context are the Fe
11	igh fficiency An application	of particular interest involves a roll formed section
12		Of particular interest is the finding that the expressio
13	al.1995a) ~ T9102RQ A1996PE	Of particular interest is that , in all informative assa
14	and the kidneys of mammals.	Of particular interest is the fact that many, probably t
15	a noise-amplifying precess.	Of particular interest is the case with which we can sim

由上述词语索引可知，*of particular interest* 这个搭配序列有两个结构变体，一个是 *of particular interest are/is*，另一个是 *is/are of particular interest*。前者是一种语法结构上的倒装形式，目的在于强调某个部分。用倒装形式时，与研究专题有关的某个具体内容置于搭配序列的右边；索引行 2、3、10、12、13、14、15 的情况即是如此。我们以索引行 2 和 3 为例来分析一下倒装形式的用法

特点。在这两个实例中，位于序列右边的 *the results obtained* 和 *the wave loads on fixed structure* 是与研究专题紧密相关的内容，因此被说成是具有 *particular interest*。而位于序列左边的一般是介绍研究领域里一般信息的词语，或者是简短提及与研究专题有关的问题的词语。索引行 2 的扩展语境如下：

2. Using knowledge gained, improvements of combustion in existing engines and its extension to high-bmep engines are discussed. Of particular interest are the results obtained from engines of up to 12(1/2) in 2. (JDEST 数据)

上述引句显示，搭配序列左边的 *knowledge*、*improvements of combustion* 和 *its extension* 等词语都是用来概略地介绍研究专题的词语。而序列右边的词语 *results obtained* 则是指研究专题某个方面具体的内容，这个具体的内容要去吸引话语社团成员的注意，引起他们的兴趣。倒装形式 *Of particular interest is/are* 不仅仅出现在句首，而且还出现在段首，如索引行 12 和 13 的情况。在这些情况下，搭配序列就像一个“联结器”，将前一句与后一句，前一段与后一段的语篇信息连接起来，使信息从一个篇章部分向另一个篇章部分有序地“流动”。

搭配序列用于非倒装形式 *is/are of particular interest* 时，研究专题或与研究专题紧密相关的内容置于序列的左边。比如，索引行 4、5 和 6 中的 *the latter*、*this case* 和 *to research further points* 就是表示与研究专题有关的内容。情况似乎是，这种形式的语篇连接功能比倒装形式要弱一些，或者说没有倒装形式明显。但是，毋庸置疑，无论那种形式，论文报道的研究专题都被声称是“特别有趣”，以引起读者重视。而搭配序列 *of particular interest* 正是实现这一语篇功能的常用词语手段之一。

4.3.2 *It is not clear*：对以往同类研究提出质疑

研究论文引语的第二个语步为"确立位置"。但该语步里的信息与第一个语步里的信息却极不相同。如果说第一个语步里的语篇信息以强调研究专题的"重要性"、"意义"、"趣味"等为其主要特点的话，在第二个语步里，语篇信息的主要特点就是强调以往同类研究存在的"问题"、"弱点"和"差距"。研究者通过指出本领域里同类研究的问题、质疑它的弱点、揭示存在的差距，从而为自己的研究建立一个空间。而证据表明，*it is not clear* 就是被用来提出质疑的一个搭配序列：

1	K+ conductance. ~ T9004R33B1994PE	It is not clear from the present results why both b
2	average print-viewing level. Also	it is not clear how Schreiber avoided the veiling i
3	les in these relationships so that	it is not clear how significant these epidemiologic
4	ecessary for the coordination and	it is not clear how to organize the specialized in
5	formably upon the Malin formation	It is not clear how long a period this erosive even
6	model proposed by other writers.	it is not clear how the approach could ever be
7	hological processing is the same,	it is not clear that the way they have internally o
8	to different conclusions. First,	it is not clear that the positive context manipulat
9	l General Esteem scale, however,	it is not clear that it was intended to be such a s
10	differential sensitivity; however,	it is not clear that an LQG design dominated by the
11	range of bulk compositions. Thus,	it is not clear to us that further modeling of the
12	for HY80 steel have been used and	it is not clear to what extent the full strength el
13	ove the quality assurance system,	it is not clear what kinds of information are nec
14	are required (of order 1 km3/yr):	It is not clear whether such recycling rates are re
15	s of English extraposed clauses,	it is not clear whether they should be considered
16	pinach (Spinacea oleracea) [442].	It is not clear whether this is a different enzy
17	se, in the absence of a fold test,	it is not clear whether the B component was acquire

上述 17 个词语索引中的 *it is not clear* 这一搭配序列都被用来质疑以往同类研究可能存在的弱点和问题。一般来说，质疑可能针对三个方面：研究方法、研究结果和研究的局限性。在 17 个

词语索引中,1、2、4、6 和 7 是针对研究方法上存在的弱点而提出问题的。由于篇幅所限,我们只以索引行 6 的扩展语境为例示之:

6. it is not clear how the approach could ever be extended to cases where the number of particles is extremely large.(JDEST 数据)

很明显,作者是在对研究方法提出质疑,认为它有缺陷,所以说不知道如何才能把这种方法用于研究粒子数目较大的情况。索引行 3、13 和 16 是针对研究结果提出质疑的,索引行 3 的完整句子是:

3. it is not clear how significant these epidemiological observations may be.(JDEST 数据)

说观察结果不知有没有显著意义,当然是对研究结果表示怀疑。索引行 10、11 和 12 是针对研究的有关局限性而提出质疑的。索引行 12 的完整句子是:

12. and it is not clear to what extent the full strength electrodes have been assessed.(JDEST 数据)

作者说不知道在多大程度上全强度电极作了评估,说明他认为研究有局限性。无论是质疑研究方法,研究结果,还是对有关局限性提出问题,都是在试图说明领域内过去开展的同类研究存在着一个“有问题的区域”(a problem area)。而这个有问题的区域正好给作者留下了研究空间。这样,作者就为自己的研究在本领域内找到了一个合适的位置。而 *it is not clear* 这个搭配序列无疑是作者用于建立研究空间的重要词语手段之一。

4.3.3 *The purpose of this paper is to* ...：宣布目的

学术论文的第三个语步是占据空间。该语步的内容与第二步紧密相承。在这里，作者要着手证明自己的声称有根有据，要回答提出的问题，要试图填补空白。其实质就是占据业已建立的研究空间。占据研究空间的方式可多种多样，但"宣布本次研究的目的"或者"宣布研究的特点"一般是大多数论文引语部分都有的一个步骤。在宣布研究目的时最常用的一个惯例化搭配序列就是 *The purpose of this paper is to*。在 JDEST 语料库中，该搭配序列有 28 个实例，下面是其中的 15 行索引：

1	t also when manufacturing them.	The purpose of the paper is to summarize those FMS
2	u	The purpose of the paper is to present some highligh
3	will be addressed. INTRODUCTION	The purpose of this paper is to preview the acute an
4	essee at Phattanooga Abstract	The purpose of this paper is to review the pioneer
5	mber of project failures.	The purpose of this paper is to explore the nature
6	5600M Eindhoven the Netherlands	The purpose of this paper is to show that modular de
7	performance. Therefore the primary	purpose of this paper is to review and integrat
8	ension of control-the negative.	The purpose of this paper is to suggest more positi
9	te computational capabilities.	The purpose of this paper is to provide a didactic e
10	island arc system. ~T4283P5	The purpose of this paper is to document field, text
11	on or environmental effects.	The purpose of this paper is to suggest that SCC pla
12	aurus are briefly described.	The purpose of this paper is to discuss the curricul
13	ratory scale or in the field.	The purpose of this paper is to examine the role of
14	ill also give such a mechanism.	The purpose of this paper is provide guidance for mo
15	arantee a high solubilization.	The purpose of this paper is to investigate the infl

这个搭配序列有几个特点值得注意。首先是 *paper* 之前限定词的使用。在 28 个实例中，只有两例用了定冠词 *the* (1,2)，其余全是代词 *this*。在研究论文中，*this paper* 和 *the present paper* 基本上同义(参见第三章 1.2.1 节)。所以，在宣布"本论文"的研究目的时，用 *this paper* 已成为确定的惯例。第二个特点是一般现在

时的大量使用。在 28 个实例中，26 例用了“is”，只有 1 例用了一般过去时形式 *was*。另外，另一例用了动词 *implies*。该句显然不属于我们讨论的情况。大量使用动词的一般现在时形式表明作者在努力强调所报道的研究与“本领域”里目前的现实具有“相关性”。第三个特点是在连系动词 *is* 之后大量使用动词不定式短语 *to* ...。除了两个例外情况，其他实例全用了动词不定式短语。从修辞角度来说，使用动词不定式似乎要更有力些。

在上述词语索引中，*paper* 是该搭配序列中主要的词汇之一。在宣布研究目的时，除了 *paper* 之外，作者还可用其他一些词，如 *article*、*investigation*、*study*、*project*。因此，*The purpose of this article / investigation / study / project / is to* 也经常出现。不过，它们远没有 *The purpose of this paper is to* ...出现的频数多；在 JDEST 语料库中上述四种情况加起来共有 21 个实例。

该搭配序列中的另外一个关键词是 *purpose*。除此之外，作者还用 *objective* 和 *aim* 这两个同义词。在 JDEST 语料库中 *objective* 类的实例（包括 *The objective of this paper*、*The objective of this study / investigation*）有 18 个，*aim* 类的实例（包括 *The aim of this paper*、*The aim of the study* 等）有 7 个。两类情况都远没有 *purpose* 的实例多。

综合上述情况，*The purpose of this paper is to* ...应当说是个典型的搭配序列，是较为标准的形式，其他各形式是它的变体。它是科技和学术研究人员宣布研究目的时最常用的词语手段。

结　束　语

本章采用定性分析与定量分析相结合的方法，研究了词语搭配的四个类别。一般搭配是普通语言里使用频率极高的习惯性词语组合，其使用与语言的地道性和自然性密切相关，是语言学习的重

要内容。修辞性搭配是基于词汇的修辞性意义的词语组合；它或多或少仍具有新颖的交际效果。判别修辞性搭配的重要手段之一是定量的统计测量标准。专业性搭配揭示了语域对词语搭配的限制作用。无论是技术词还是次技术词都吸引一定数量的典型搭配伙伴，以实现一定的概念意义或语篇意义。惯例化搭配序列显示了词语组合的高度因循性。无论是普通语言还是专业领域里的语言使用都有其典型的惯例化词语组合序列。研究专业领域里的惯例化搭配序列，探讨其实现的意义与功能尤为重要。本章仅仅是对区分各类搭配作了一些尝试性的努力，以期使研究者和学习者对不同类别的搭配有一种概念上的认识，更有利于研究和学习。然而，各类词语搭配间并没有不可逾越的鸿沟。相反，它们之间有着必然的联系。本章只是就搭配的主要特征将它们作个粗线条的轮廓分析，并不是要将它们截然切分开来。



第七章 语料库语言学与学术英语语体研究概述

1 学术英语语体研究的重要性

任何语篇都是在一定的语境下产生的。相同或相似的语境下使用的语言有许多共同特征。这些特征的总和定义了该语境下所用语言的语体,并使这一语体明显地不同于其他语境中使用的语言的语体。ESP、EAP 研究人员已对不同语域做过大量研究,并普遍认为不同的文类在各语言学层面都存在重要的、系统的差异。语言使用者在不同的场合生成和需要理解的语篇是有语体差异的。说话人/作者要说出或写出地道、贴切的英语语篇表达自己的思想,听话人/读者要正确地、最大限度地接受和理解语篇中提供的信息,都必须对不同语境下的英语语体有所了解。可见,对英语语体的研究和学习是很有现实意义的。

乔丹(R.Jordan 1997)做过一项研究,调查一所英国大学上写作课的硕士留学生在写作方面的困难。每类困难从"不难"到"很难"分成六等。各项中,选择"难"和"很难"的老师和学生的百分比分别为:

1. 学生	
vocabulary	62%
style	53%
spelling	41%
grammar	38%

2. 老师	
style	92%
grammar	77%
vocabulary	70%
handwriting	31%

punctuation	18%	punctuation	23%
handwriting	12%	spelling	23%

由此可见，排序虽有先后，但英语的语体，不论对学生还是对老师，都是一个难题。

一种语言的语体风格是千变万化和极其复杂的，英语也不例外。本族语者，尤其是受过教育的本族语者通常能够根据语境本能地选择适当的母语语体，并随着语境的变化而变换语体。许多学习英语的中国学生在使用英语进行交际时出现障碍，原因之一是不能根据语境使用得体的语言。要掌握地道的英语、具备用英语顺利地完成各种交际的能力，英语学习者对与自己密切相关的文类的语体风格的学习这一关是躲不过的。只有对一种语体的特点、该语体与其他语体的异同有所了解，才能在学习过程中根据这些特点和差异去掌握这种语体，并在交际过程中根据语境的需要恰当地选择或运用。学习者有效地运用语言以及追求语言的得体性是语言生成能力(包括说、写、译等能力)的重要组成部分。

在科学知识迅速国际化的今天，对学术交流模式的了解和认识是很迫切的。对语言知识的一知半解可能影响学者间的交流，并最终对学术合作和进步造成不利影响。斯韦尔斯(J. Swales 1990) 对英国一所大学的图书馆中关于健康和经济学方面的杂志中的论文进行随机抽样，统计其中非英语本族语作者的比例。他发现，只有20%的文章源自非英语国家。这些文章的作者分布在117个地点，而这些地点中只有21个在第三世界。也就是说，在国际专业杂志上发表的论文中来自第三世界科学家的不足4%，这当然远不能反映第三世界科学家对科学的贡献，斯韦尔斯认为这是语言障碍造成的。非英语本族语者要在国际学术界得到承认，需在学术英语方面下很多工夫。语言在知识的传播过程中起着举足轻重的作用，尤其是英语，在学术界拥有国际语言的地位，是其他语种学者进入国际学术界的重要途径。因此，近来对学术英语课程的要求日益增加，学术英语的研究也日益受到重视。在我国，学者们同样要了解本专

业在国际上的发展状况，或报告自己的发现或发明，以取得国际学术界的认同。我国大学英语教学的主要目标是使学生能以英语为工具进行学术交流。因此，在大学英语教学大纲中对学术英语阅读课的设立有专门的规定、对写作能力的培养也有一定的要求。可见，我们有必要对学术英语的语体风格作深入、细致的探索。

语体研究是语言学研究的重要一环。它能帮助我们了解同种语言中不同语言品种的语言学特征、理解其使用上的限制，从而深入了解语言与社会、人与环境的关系。语体研究是极富实用价值的。除了能使语言使用者更加灵活、有效地运用语言外，由于其注重文本特征的研究和描述，还常被用来确定文学作品的归属，确定作品的写作时间等。语体研究还曾用于解决法律问题，如确认合同、遗嘱或匿名信的作者等。在语料库建立方面，语体研究也有重要意义。它有助于人们选择有代表性的语料并对语料进行科学的分类，因而更有利于开展研究工作。语体知识有助于教师分析讲解课文及指导学生如何在真实情况下使用语言。为方便学习者对英语语体的掌握，必须进行语体的研究，描述不同语境下所用语言的语体风格，以便学习者少走弯路。针对大学英语教学，开展对学术英语的研究，特别是其语体的研究，有着重大的现实意义。

2　语体风格的量化研究

人们在使用语言的过程中对语言项目使用频率高低会留下印象，在记忆词汇、语法结构时，会同时储存这些词汇、语法项目在语言中出现的概率。例如，懂英语的人不仅认识 *begin* 和 *commence* 这两个词，还自然地知道 *begin* 的使用频率远比 *commence* 高。人在使用语言时会不自觉地根据语境参考潜意识中的概率知识，选择恰当的表达方式。语言使用者头脑中之所以会有对语言项目使用概率的经验积累，是因为语言项目的使用频率

确实存在差别。从一定意义上讲,语体是不同文本中各语言单位的使用频率的差异引起的。很多语言学家承认语体受语言单位使用频率的影响。

人们可以用不同的方式表达思想。不同文本中各种表达方式使用频率上的差异导致了不同语体风格的存在。我们对某文本的语体风格的印象是由于该文本中的语言项目的密度与这些项目在语境相关的对照文本或文本体系中的密度明显不同造成的。描述语体风格即是描述一个文本或文类在多大程度上偏重于某种表达方式。这种描述一般是通过对比不同文本或文类来进行的。只有通过比较才能发现语言单位在不同文本中的用法及使用频率上的区别。对一般人来讲,这种比较建立在对语言的以往经历的基础上,是靠直觉和印象的。例如,有经验的读者会对一个文本的语体风格有这样或那样的感觉。但是要他明确讲出是什么因素造成这种印象则不太容易。语体研究人员对某种语体风格的判定一般要有数据支持,否则得出的结论将是不客观的。

语体的属性决定了语体研究的量化特点。语体量化研究的例子是很多的。梅里克(L. Millic)(Oakes 1998: 215)对斯威夫特(Swift)的语体风格做过一些量化的研究。他将斯威夫特的作品与其他一些作者的作品进行对照,发现斯威夫特与其他作者不同,其作品中有较多的非限定性动词(不定式、分词和动名词)、引导连接词(introductory connectives, 包括并列、从属连接词及连接副词)及由不同词类构成的三词结构(如介词 + 限定词 + 名词)。

对两个文本进行比较时,我们可能发现一个语言项目出现在一个文本中,而不会出现在另一个文本中,可能在一篇中比在另一篇中更频繁,也可能在两篇中频率大致相同。如果某一项目的密度在两个文本中不同,该项目即可作为区分两个文本的重要语体特征(即语体标记)之一。如果某项目的密度在两个文本中大致相同,那么,对这两个文本来讲,它们是非标记性的,或中性的。必须注意的是,做比较的文本不能是任意的。充当标准的文本必须在语境上与

待比较文本有密切的关系。通过对语境上相关的文本的对比而确定的语体风格特征才是有意义的。如，温特(W. Winter)(Enkvist 1973)调查了来自四种相关文类的30个德文文本、共60 000个句子中句首词的词类、从句长度及句子的复杂性。他发现所调查的语言单位的使用频率和语境有显著的相关性。以副词开始的句子在书面语中密度较高，特别是在科技文章和散文中。而在戏剧中的密度则较低。限定性动词在科技文章中的密度低，在戏剧和小说中较高。这就是说，以副词开始的句子及限定性动词是区分科技文章及戏剧的语体标记。恩克韦斯特(Enkvist 1973)认为这些发现可以解释成语境概率(contextual probabilities)。根据这一概率知识，既可以从语境推知语言项目的密度，也可以从语言项目的密度推知语境的种类。他说，"一旦语体标记和语境范围的对应关系建立起来，其作用将是双向的：如果我们了解语体标记，我们就可得出关于语境的结论；如果我们了解语境，我们便可以预测何种语体标记将出现于其中"(1973：96)。

在对语体标记使用频率的差异及差异的显著性进行量化分析时，统计方法起很重要的作用。统计学用于语体研究有很长的历史，据恩克韦斯特(1973：129)所说，19世纪中叶，统计学最早被用于语体研究。齐波夫(Zipf)(见Enkvist 1973)研究了词频和词长的关系，发现不论在中文、拉丁文还是英文中，词长与词的使用频率趋向于成反比关系。他还发现词在频率词表中的排列序号与词的频率之积趋向于一个常量。不过这一关系对词频表中部的词来说是适用的，而对两端的词，即最常用及最罕用的词则不太准确。以前，为了进行语体的量化分析，很多研究把注意力集中在容易定义、容易量化且统计起来比较方便的语言项目上面，如最常用到的项目是词长及句长、词在句中的位置、词汇的选择及使用频率等。统计学理论及计算机技术的发展不仅使这类量化研究更方便，而且使从前无法做到的统计工作得以顺利进行。

我国的英语语体研究大多是定性的描述，而且至今仍停留在这个层面上。为了对学术英语语体作出较为详尽的、客观的、系统的

分析描述，本文拟通过对从学术英语语料库 JDEST 和一般英语语料库 LOB 中随机选取的语料进行定量分析，用客观数据来反映学术英语的语体风格。并对一般英语和学术英语的语体风格的差异进行较系统、客观的分析比较。

3 语料库与语体学研究

3.1 语料库与语体学研究

语言项目的概率积累与语言本身一样是很难通过内省的方式分析、研究的。一方面由于经验的限制，人们对语言的直觉体验与语言事实可能有偏差。例如，奥依(Ooi, V. 1998:50)根据自己的经验认为"手机"在英文中是 *hand phone*，但他惊异地发现 COBUILD 等词典中都未收录 *hand phone* 一词。它们收录的是 *mobile phone* 或 *cellular phone*。于是，他检索了包含 3.2 亿词的 COBUILD 语料库，发现 *mobile phone* 用了 2479 次，*cellular phone* 用了 447 次。*hand phone* 只出现了 1 次，而且是出现于取自澳大利亚报刊的语料中。这说明受过教育的英、美人是不用 *hand phone* 这个词来指"手机"的。作者之所以有原来的印象，是由于该词在作者生活的环境中被普遍使用。可见，频率上的特征靠直觉是比较难以捕捉的。

语体研究有量化的、对比的特点。因此要研究某文本的语体风格，最好将该文本与某一适当标准进行对比，来分析其中语言项目的使用频率与所选标准的差异，从而认识该目标文本相对于所选标准的一些特点，并以此指导我们对该文本的语体风格的学习和掌握。对语料库中的语料进行统计是发现语言项目出现概率的有效方法。语料库一般是语言品种或语言的标准样本，因此语料库是比较一个作者、一个或一些文本与同种语言其他特殊品种异

同的很有价值的基础。语料库在分析文本或文类间的概率差异方面作用尤其显著。1964 年 BROWN 语料库问世后，基于语料库的英语语体研究就开始了。最初，这类研究局限于 BROWN 语料库中所包含的 15 个美国英语文类之间的比较。由于 LOB 语料库与 BROWN 语料库的结构类似，而 LOB 的语料来自英国英语，可以通过两者的对比进行地域变体的研究。而 LLC 口语语料库的建立为探讨口语与书面语的异同提供了条件。HELSINKI 语料库则是研究历时性(diachronic)变体的好帮手。语料库还被用于研究个人的文体风格。如梅里克(McEnery & Wilson 1996: 101)建立了80 000词的语料库。样本取自 1675—1752 年英国出版的散文。这些材料代表当时受过教育的人可能涉猎的广大阅读范围。建库的目的是要将斯威夫特的文体风格与当时书面英语的标准样本进行比较。

当今语料库最重要的特征是机读性。机读语料便于检索，且有利于增补新的语言材料。它提供的语料更有真实性，因而更有利于对语言的不同层面的描写研究，更有利于进行不同语体的比较，是开展大范围语言量化研究的极好的资料来源。世界上已经有不少规模较大的语料库，研究者个人也开始建立适合于自己研究的小型语料库。这些语料库在语体学研究中发挥着重要的作用。

3.2 JDEST 语料库与学术英语语体研究

大学英语教学的目的主要是培养学生在学术领域使用英语的能力。为了配合大学英语教学及进行学术英语的研究，上海交通大学创建了学术英语语料库 JDEST。该语料库于 1983 年始建，经过增补，现有文、理、工、医等 4 大类，33 个专业，超过 400 万词，是一个初具规模的“学术英语语料库”。语料是严格按照随机抽样原理，从英语国家正式出版物中选取的。每个专业的语料由约 500 个词的连续语篇组成。选材时尽量保持了语料的完整性。JDEST 语料库建成后对理工科大学英语教学大纲词汇表的制定、教材编写、常用

词词表的制定及学术英语特征的研究起到了积极作用。利用JDEST语料库的统计结果确定大学英语教学大纲词表,使得大学英语教学词汇的数量、范围的确定有了科学的依据。JDEST语料库对统计学术英语的常用词及研究学术英语的词汇特征也有一定帮助。通过观察JDEST的频率词表,并将其与其他语料库的词频表对照,可以发现很多凭直觉难以注意到的学术英语词汇频率方面的重要特征。比如,通过对比JDEST及一般英语语料库LOB的频率词表,我们很容易看出两库中最常用的动词和名词的差异。(见表7.1。)也可以发现词的使用情况,比如,同一词根的不同词形在不同语域的使用频率是不同的。(见表7.2。)

这些数据都从某种意义上反映了学术英语不同于一般英语的语体特征。JDEST语料库是进行与学术英语有关的各项研究工作的得力助手,特别是在学术英语语体研究方面,有尤其重要的意义。由于JDEST语料库只限于学术英语,400万词的规模应该说有一定代表性,能为我们研究和描述学术英语的语体特征提供可靠的、可量化的、有代表性的数据。

表7.1 一般英语及学术英语中最常用的动词和名词

动词		名词	
一般英语	学术英语	一般英语	学术英语
say	use	year	system
make	make	man	time
word	show	time	process
see	control	people	result
know	increase	way	problem
get	work	life	material
go	form	part	case
come	give	day	effect
use	change	world	surface
take	design	course	point

表 7.2 不同词形在两库中的不同使用频率

词 形	频 率		次 序
	JDEST	LOB	
active	678	92	JDEST: active>activity>actively
actively	7	2	
activity	472	114	LOB: activity>active>actively
abdomen	20	8	JDEST: abdominal>abdomen
abdominal	80	15	LOB: abdominal>abdomen
abrupt	34	7	JDEST: abrupt > abruptly > abruptness
abruptly	27	24	
abruptness	3	1	LOB: abruptly>abrupt>abruptness

作者从 JDEST 语料库 33 个专业领域的文本中随机抽取了约 550 000 词的学术英语语料，从一般英语语料库 LOB 各文类中随机抽取了共约 193 000 的一般英语语料，并调查了一些语法项目在所选样本中出现的次数，以比较学术英语与一般书面英语的语体上的差别。为了避免样本长度差异对数据的影响，所得的原始数据均转化成了每 1000 词中出现的次数。表 7.3 给出了所调查的项目在学术英语和一般英语中频率较高的语法项目。在学术英语与一般英语中使用频率无显著区别的项目见表 7.4。频率有显著区别的项目见表 7.5。

表 7.3 学术英语和一般英语中频率最高的 11 个语言项目

学术英语		一般英语	
语言项目	频率	语言项目	频率
名词(nouns)	264	名词(nouns)	224.45
介词(prepositions)	132.55	介词(prepositions)	114.43

（续表）

学术英语		一般英语	
语言项目	频率	语言项目	频率
作定语的形容词(attributive adj.)	79.86	作定语的形容词(attributive adj.)	49.55
现在时(present tense)	45.62	现在时(present tense)	40.16
名词化(nominalizations)	32.67	过去时(past tense)	26.23
副词(adverbs)	30.73	第三人称代词(third person pronouns)	24.93
BE 作动词(BE as main verb)	19.91	副词(adverbs)	18.34
无 by 被动态(agentless passives)	17.24	BE 作动词(BE as main verb)	18.33
过去时(past tense)	14.84	名词化(nominalizations)	15.75
不定式(infinitives)	12.82	不定式(infinitives)	14.77
作表语的形容词(predicative adjectives)	10.78	第一人称代词(first person pronouns)	12.04

表 7.4　无显著区别的语言项目（X 表示差异不显著）

语言项目	频率		差异显著性
	学术英语	一般英语	
名词(nouns)	264	224.45	X
介词(prepositions)	132.55	114.43	X
BE 作动词(BE as main verb)	19.91	18.33	X
完成体(perfect aspect verbs)	4.00	4.08	X
现在分词短语后置定语(present participial WHIZ deletions)	2.14	1.65	X

（续表）

语言项目	频率		差异显著性
	学术英语	一般英语	
条件从句(adv. sub. — condition)	2.00	1.87	X
让步从句(adv. sub. — concession)	0.93	0.75	X
过去分词短语作状语(past participial clauses)	0.45	0.44	X
特殊疑问句(WH question)	0.23	0.39	X

对语料库中的语料进行频率统计可以发现语料的诸多特征，此处仅列举了一些语言项目使用频率在不同文类中的差异，以供参考。学术英语相对于一般英语的这些数据特征有助于我们认识和描述学术英语，有助于对学术英语的理解和进行学术英语写作，也有助于学术英语教学。不通过对真实语料的量化分析，仅凭直觉，是不可能发现这些频率特征的。但是，可以看出这些特征比较零散，缺乏系统性，不能使我们对某一文类语体风格形成较全面的认识。MD/MF 模型为发现语体标记使用上较为系统的特征提供了有力的工具。

表 7.5 有显著区别的语言项目(* * P<0.01，* P<0.05)

语 言 项 目	频 率		差异显著性
	学术英语	一般英语	
作定语的形容词(attributive adj.)	79.86	49.55	* *
现在时(present tense)	45.62	40.16	* *
名词化(nominalizations)	32.67	15.75	* *

（续表）

语言项目	频　率		差异显著性
	学术英语	一般英语	
副词 (adverbs)	30.73	18.34	* *
无 by 被动态(agentless passives)	17.24	9.95	* *
过去时(past tense)	14.84	26.23	* *
不定式(infinitives)	12.82	14.77	* *
作表语的形容词(predicative adjectives)	10.78	6.18	* *
第三人称代词(third person pronouns)	8.11	24.93	* *
表示个人感觉的动词(private verbs, e.g., anticipate, believe, feel)	7.43	9.29	* *
动名词(gerunds)	6.56	8.38	* *
代词 IT(pronoun IT)	6.03	8.61	* *
过去分词短语作后置定语(past participial WHIZ deletions)	5.20	2.73	* *
第一人称代词(first person pronouns)	3.86	12.04	* *
连接词(conjuncts)	3.80	2.43	* *
有 BY 被动态(by passives)	3.17	1.51	* *
时间状语(time adverbials)	3.07	3.58	* *
地点状语(place adverbials)	2.81	0.72	* *
THAT 引导的宾语从句 (THAT verb complements)	2.53	3.51	* *

（续表）

语言项目	频率		差异显著性
	学术英语	一般英语	
现在分词从句（present participial clauses）	1.99	2.70	* *
表示公开行为的动词（public verbs, e.g., assert, complain, mention）	1.67	3.44	* *
存在句（existential THERE）	1.64	2.10	* *
第二人称代词（second person pronouns）	1.07	2.55	* *
原因状语从句（adv. subordinator — cause）	0.77	0.40	* *
指示代词（demonstrative pronouns）	0.50	3.77	* *
不定代词（indefinite pronouns）	0.50	1.71	* *
指示词（demonstratives）	0.37	5.15	* *
缩略形式（contractions）	0.36	1.53	* *
省略 THAT 的定语从句（THAT deletion）	0.22	0.69	* *
DO 作代动词（DO as pro-verb）	0.12	0.39	* *
分裂不定式（split infinitive）	0.06	0.00	* *
THAT 从句作形容词的补语（THAT as adj. complements）	0.03	0.12	* *

4　MD/MF 模型

4.1.　基本思想

MD/MF 是多维度（multi-dimensional）、多特征（multi-

feature)的缩略形式。多特征指模型建立过程中利用了大量的语言项目。语言的语体风格从非正式到正式，或从口语化到书面语化等的变化是一个连续体，而非孤立的两极。这一连续体在该模型中被称为维度。这一模型是贝博(D. Biber 1988)在研究口语和书面语的特征时建立的。这项研究证明语域之间存在系统的差异，这些差异仅从某个维度是不足以描述清楚的，需通过不同的维度来分析。

该模型的基本思想是根据所要研究的对象，确定大量可能有区别意义的语言项目，调查其在研究对象中的使用概率，再用数理统计的方法确定这些语言项目的相关程度，即确定哪些语言项目在文本中是同现还是相互排斥。相关程度高的项目属于同一维度。然后再从功能的角度，根据各维度所包含的语言项目及这些项目在研究对象中出现的频率，对研究对象在这些维度上的特征进行分析并对其语体特征及其间相互关系进行描述。该模型从多个维度上对某文本的特征进行描述。每个维度都包含一组在篇章中频繁共现或互斥的语言项目。其特点在于它强调从多个维度描述语体差异，强调维度的连续性，而且维度是通过经验的方法而非想当然地确定的。模型注重语体间整体的异同，而非在个别语言项目上的差异，从而避免了以往一次只研究一个语言项目的那种做法。

4.2 简例

贝博(1988) 用了一个简单的例子说明 MD/MF 模型的原理。为了使读者更好地理解这个模型，简单介绍一下这个简例。对话和科技文章可从普通/专业化、有准备/无准备等不同角度来观察其差异。对话是很常见的语体，科技文章一般人则不太会涉猎；对话一般不会预先准备，科技文章的写作则相反。当然，在只考虑这两种语体时，我们似乎是在讨论两种孤立的语体。但当加入诸如小组讨论这一语体时，语体变化的连续性就会体现出来。因为小组讨论在专业化与否、有准备与否方面是居于对话和科技文章之间的。这样

我们就可以从是否专业化及是否有准备这两个维度上来描述对话与科技文章的异同。而这两个维度各自包含一系列语言项目，或者说由一系列反映文本是否有准备、是否专业化的语言项目来定义。为了说明模型建立的理念，贝博从这3种文类中各选了一段，统计了4项语法项目，即被动态、名词化、第一、二人称代词及缩略形式在其中的频率，结果如下：

表7.6　4项语法项目在3个语篇中的频数

	被动态	名词化	第一、二人称代词	缩略形式
对话	0 (0)	1 (0.48)	12 (10.2)	6 (5.1)
科技文章	3 (6.8)	5 (11.4)	0 (0)	0 (0)
小组讨论	2 (2.2)	4 (4.3)	10 (10.8)	3 (3.2)

表中括号外的数字是各语法项目出现的总次数，括号中的数字是每100词中该项目出现的次数。因为所选文本长度不同，为使数据具有可比性，必须将数据换算成在同样词数的文本中语言项目出现的频数。

从表中的统计结果来看，在对话中被动语态用得少，名词化也用得少；而科技文章中被动态用得多，名词化也用得多。也就是说，被动态和名词化现象是同现的，即属于同一维度。同样道理，第一、二人称代词和缩略形式在对话中都高，而在科技文章中都低，故属于同一维度。小组讨论中四项频率都较高，说明必须用被动态/名词化和第一、二人称/缩略形式两个维度来描述这三种语体，而不能把这四项视为一个维度。

这种语言项目的共现现象应该能从功能的角度进行解释。对于这个简单的示例来说，被动态与名词化共现是因为它们都为书面形式，有正式色彩。在此处，它们频繁出现在一般说来比较正式的科技文章中，因此该维度可定义为潜在的信息性、抽象性维度的具体反映。定义了这样一个维度后，对具体篇章或语域来说，我们可

以计算该维度所包含的语言项目在其中出现的频率，求其在该维度上的得分，从而确定其在该维度上的位置、其语体特征及其与其他篇章或语域的关系。

求出的各语体的得分可以描述在坐标上，以便清晰地看出语体之间的关系。该简例中三种语体分布应该类似图 7.1。该图表明，科技文章和小组讨论在被动态/名词化维度上有共同特点，而在第一、二人称代词/缩略形式维度上，小组讨论与对话很相似。

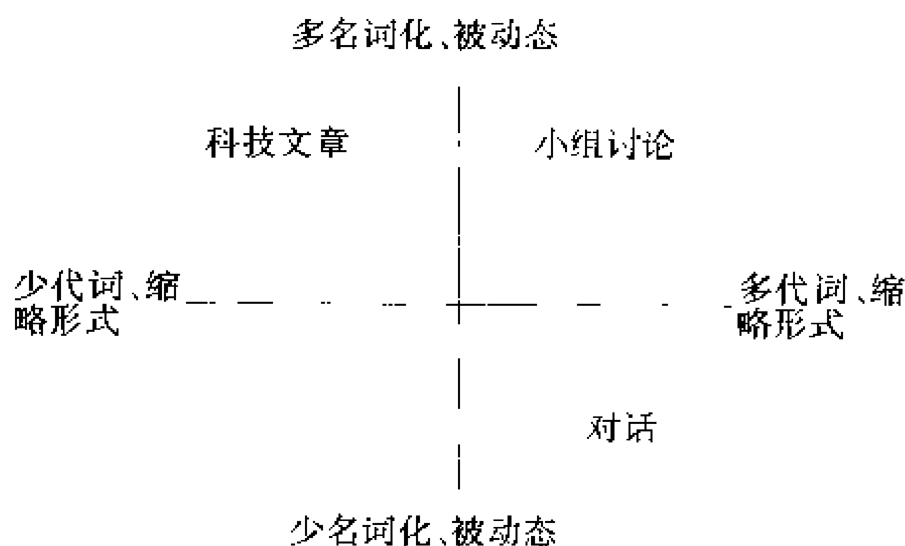


图 7.1　语篇的分布

4.3　实际模型的建立

4.3.1　抽样及语言项目的确定

该模型的建立要求所选语料能够代表所有可能的语言变体，所选语言项目能够反映所有可能的共现规律，即所选语料及语言项目都必须具有很好的代表性。贝博等人从 LOB 语料库及 London-Lund 语料库中选取了 481 个(约 960 000 词)文本。他们对以往的语体学研究进行了分析，选择了这些研究中的 67 项语言项目，如介词短语、时态、各类代词及名词化现象等等作为分析对象。

4.3.2　频率统计

然后对这67项语言项目在481个文本中的使用频率进行统计。统计过程中充分利用了计算机技术快捷、准确的优势，大大节约了人力、物力和时间。但是，有些语言项目的统计仍旧不得不靠人工整理。每个项目在每一篇章中的频率都换算成每1000词出现的频率。

4.3.3　因素分析

因素分析的目的是要通过数学的方法确定这67个语言项目的共现规律。通过因素分析可以将大量的原始数据减少为一组因素。每个因素代表原始数据中能够归纳在一起的那部分信息。在此，每个因素代表一组共现频率很高的语言项目。因素由相关矩阵中抽取，通过语言项目的频率间的相关系数决定。相关系数高的语言项目共现频率高，属于同一因素。在贝博的研究中，因素分析结果生成了六个主要因素。表7.7列出了第一和第五因素包含的语言项目。以第五因素为例，其中权重为正的语言项目有共现于文本的趋势，而当某文本中有这些因素时，权重为负的项目往往不会出现或出现频率极低。括号内的项目在其他因素上有更高的权重，为了确保因素间彼此独立，在计算因素分数时，这些项目在此因素上不予考虑。

4.3.4　计算维度分数并对各维度进行阐释

维度分数是文本在某一维度上的得分。计算维度分数之前，要将所有项目的频率换算成均值为0、标准差为1的Z分数。这一过程能够减少那些频率非常高的项目对维度分数的不良影响。接下来求每个文本的维度分数，以便在各维度上对这些文本进行对比。一个文本在某一维度上的得分为该维度所包含的权重为正的语言项目在该文本每千词中出现频率的Z分数之和减去权重为负的语言项目的Z分数之和。

最后从功能角度对这些维度进行阐释。语言项目之所以在文

本或语域中同时出现，是因为它们有共同的潜在的交际功能。比如，构成第二因素的语言项目（见表 7.8）中，过去时态动词（past tense verbs）用于描述过去的事件，第三人称代词（third person pronouns）用于指代故事中涉及的人物，完成体动词（perfect aspect verbs）表示过去发生但有持续性影响的动作，表示公开行为的动词（public verbs）用于表示间接引语，含否定词 no 的否定（synthetic negation）更具文学色彩，而现在分词词组（present participial clauses）多用来使描写更为生动。这些项目都是叙述性的。因此该因素可解释成叙述性—非叙述性维度。贝博对 6 个维度都做了细致的分析并分别做了定义，它们分别为“交互性/信息生成”、“叙述性/非叙述性”、“所指明确/所指依赖环境”、“劝诱/非劝诱”、“抽象/非抽象”和“即时性信息详述”等维度。每个维度都包含一组独特的语言项目、都定义了口语及书面语差异的一个方面、背后都有一种独特的语言功能。

由于建立时调查了大量语料，其中包括多种语体形式，且调查的语言项目数量也相当可观，MD/MF 模型实际是具有普遍意义的。该模型曾在口语/书面语以外的其他研究中被多次采用，并取得成功，证明其对同类的对比研究来说是一种行之有效的模型。

表 7.7　MD/MF 模型中的两个维度及其包含的语言项目

FACTOR 1	
private verbs (e.g. believe, know)	.96
THAT deletion	.91
contractions	.90
present tense verbs	.86
2nd person pronouns	.86
DO as pro-verb	.82
analytic negation (e.g. That's not likely.)	.78
demonstrative pronouns	.76
general emphatics (e.g. a lot, for sure)	.74

（续表）

FACTOR 1	
1st person pronouns	.74
pronoun IT	.71
BE as main verb	.71
causative subordination (e.g. because)	.66
discourse particles (e.g. *sentence initial* well, now)	.66
indefinite pronouns	.62
general hedges (e.g. something like, almost)	.58
amplifiers (e.g. absolutely, extremely)	.56
sentence relatives	.55
WH questions	.52
possibility modals	.50
non-phrasal coordination (*clause initial* and)	.48
WH clauses	.47
final prepositions	.43
adverbs	.42
conditional subordination	.32
nouns	−.80
word length	−.58
prepositions	−.54
type/token ratio	−.54
attributive adjectives	−.47
place adverbials	−.42
agentless passives	−.39
past participial WHIZ deletions (e.g. things done by ...)	−.38
present participial WHIZ deletions (e.g. events causing ...)	−.32

FACTOR 5	
conjuncts (e.g. consequently, furthermore)	.48
agentless passives	.48
past participial clauses (Built in a single week, the house ...)	.42
BY-passives	.41
past participial WHIZ deletions	.40

（续表）

FACTOR 5	
other adverbial subordinators (those other than causative, concessive and conditional adverbial)	.39
predicative adjectives	.31
type/token ratio	−.31

表 7.8 第二个维度所包含的语言项目

FACTOR 2	
past tense verbs	.90
third person pronouns	.73
perfect aspect verbs	.48
public verbs (e.g. assert, declare, say)	.43
synthetic negation (e.g. no answer is ...)	.40
present participial clauses	.39
present tense verbs	−.47
attributive adjectives	−.41
past participial WHIZ deletions	−.34
word length	−.31

5 学术英语的语体特征

作者从JDEST语料库及LOB语料库中抽取了部分语料，利用贝博的MD/MF模型对其进行分析，以描述学术英语语体及其他一些英语书面语语体的差异。这些语料共分23个文类，其中9个来自JDEST语料库，分别为：专著、科普读物、手册、摘要、专利、教科书、书评、技术报告和学术论文。14个来自LOB语料库，分别为：新闻报道、社论、新闻述评、宗教、技能和爱好、民间传说、自传、公文、一般小说、侦探小说、科幻小说、冒险小说、爱情小说和幽默。样本的组成如

表7.9。作者调查了贝博所定义的各维度所包含的语言项目在该样本中的频率，将其换算成1000词中的出现频率，再求其Z分数，最后求出所有文类在6个维度上的因素分数，见表7.10。

表7.9　样本的组成

来自JDEST语料库的语料		来自LOB语料库的语料	
文　类	词　数	文　类	词　数
专著(SL)	15 553	新闻报道(LUA)	19 468
科普读物(SP)	14 767	社论(LUB)	20 157
手册(MA)	16 071	新闻述评(LUC)	20 472
摘要(ABS)	17 093	宗教(LUD)	24 291
专利(PA)	6918	技能和爱好(LUE)	22 757
教科书(TE)	16 327	民间传说(LUF)	22 284
书评(BR)	10 428	自传(LUG)	23 770
技术报告(TR)	11 755	公文(LUH)	21 846
学术论文(TH)	33 381	一般小说(GF)	14 733
		侦探小说(DF)	14 733
		科幻小说(SF)	14 811
		冒险小说(WF)	12 104
		爱情小说(RF)	14 146
		幽默(LUR)	18 133
总词数	131 438	总词数	263 705

现选维度1(D1)及维度5(D5)进行分析。表7.7中Factor 1和Factor 5给出的分别是因素分析得出的第一、第五维度所包含的语言项目。贝博将D1定义为“交互性/信息生成”维度，D5为“抽象/非抽象语体”。这是因为从功能的角度讲，D1包含的权重为正的语言项目(以下简称正特征)多用于面对面交流、个人色彩较浓的场合，权重为负的语言项目(以下简称负特征)则多用于注重信息

生成的场合;D5 中没有负特征,其正特征多出现于抽象与正式的场合。在维度 D1 上因素分数高的文类使用该维度中的正特征较多,负特征较少,因而交互性较强,因素分数低的文类中则负特征较多,正特征较少,故信息性较强。图 7.2 给出的是 23 个文类在这个两维空间内的分布情况。

图 7.2 中各文类的分布反映了列在表 7.7 中的组成维度 1 和 5 的语言项目在这些文类中出现的频率的相对差异。例如,学术英语在 D1 上一般有很大的负分数(在 -25.39 和 -1.12 之间)。这说明名词、长词及介词等(D1 上的负特征)在学术英语中的出现频率很高;相反,表示个人感觉的动词(private verbs),省略 that 的定语从句,缩略形式等(D1 上的正特征)出现频率很低。学术英语在 D5 上有较大的正因素分数(在 1.15 和 7.85 之间)。这反映出学术英语中会频繁出现连接词、无 by 被动态、过去分词短语、有 by 被动态等(D5 上的正特征)。

表 7.10　23 个文类在 6 个维度上的因素分数

	D 1	D 2	D 3	D 4	D 5	D 6
ABS	-25.39	-3.69	5.34	-4.39	1.15	-0.08
BR	-19.07	-2.7	7.39	-5.87	4.3	2.13
MA	-1.12	-4.91	1.37	1.47	5.55	-1.59
PA	-23.76	-5.67	3.82	-4.99	6.53	-4.4
SL	-10.07	-4.5	1.81	2.59	3.01	-0.67
SP	-3.3	-3.42	0.36	-3.54	1.93	-0.39
TE	-2.05	-4.82	1.59	-5.42	7.85	-3.3
TH	-9.9	-3.23	0.84	-3.94	6.82	0.33
TR	-16.18	-3.51	0.57	-1.44	2.5	-0.57
LUA	-11.89	2.09	-0.18	1.45	-3.55	1.51

（续表）

	D 1	D 2	D 3	D 4	D 5	D 6
LUB	3.9	2.16	1.45	5.24	−2.09	2.03
LUC	−0.06	−0.1	3.09	−3.04	−2.96	−0.51
LUD	9.29	−0.59	4.75	2.76	−0.77	2.18
LUE	4.41	−1.79	−1.09	1.95	−1.17	2.71
LUF	2.35	−0.59	2.92	3.69	−0.72	−0.47
LUG	−1.05	0.53	3.7	−0.27	−1.71	−0.74
LUH	−15.04	−1.79	5.48	2.69	5.08	−0.65
LUR	10.86	0.79	−2.07	−0.39	−0.19	−0.96
DF	30.58	7.61	−9.13	1.54	−7.45	2.73
GF	18.07	6.39	−5.66	2.2	−6.21	0.16
RF	28.87	6.8	−10.08	5.9	−8.53	2.66
SF	11.49	7.15	−5.95	1.85	−4.36	−2.84
WF	19.17	7.81	−10.32	0.01	−5.05	0.71

在另一端，爱情小说和侦探小说在D1上有最大的正分数，说明该维度上的正特征（表示私下行为的动词、缩略形式等）在其中出现频率很高，而D1上的负特征（名词、长词等）在其中出现的频率则很低。爱情小说和侦探小说在D5上有很高的负分数，说明其中词出现较少，也没有by被动态等项目。

从图7.2可见，这23个文类即使在这两维空间中也有很显著的差别。学术英语分布在右下方，小说分布在左上方，而其他文类分布在二者之间，表明它们之间是有系统差异的。与其他文类不同，学术英语信息性极强（D1）且极其抽象（D5）。另一方面，小说表现出极强的交互性及非抽象性的语言学特征。民间传说、自传等文类则在抽象性和信息性方面都处于中等水平。图7.2显示学术英语包含的各文类在坐标图上位置不同，说明学术英语内各文类之间也有系统性的差异性（但较小）。如专利和摘要的信息性比手册、教材及科幻小说强，而交互性则不如后者。一般英语也有类似的特点。

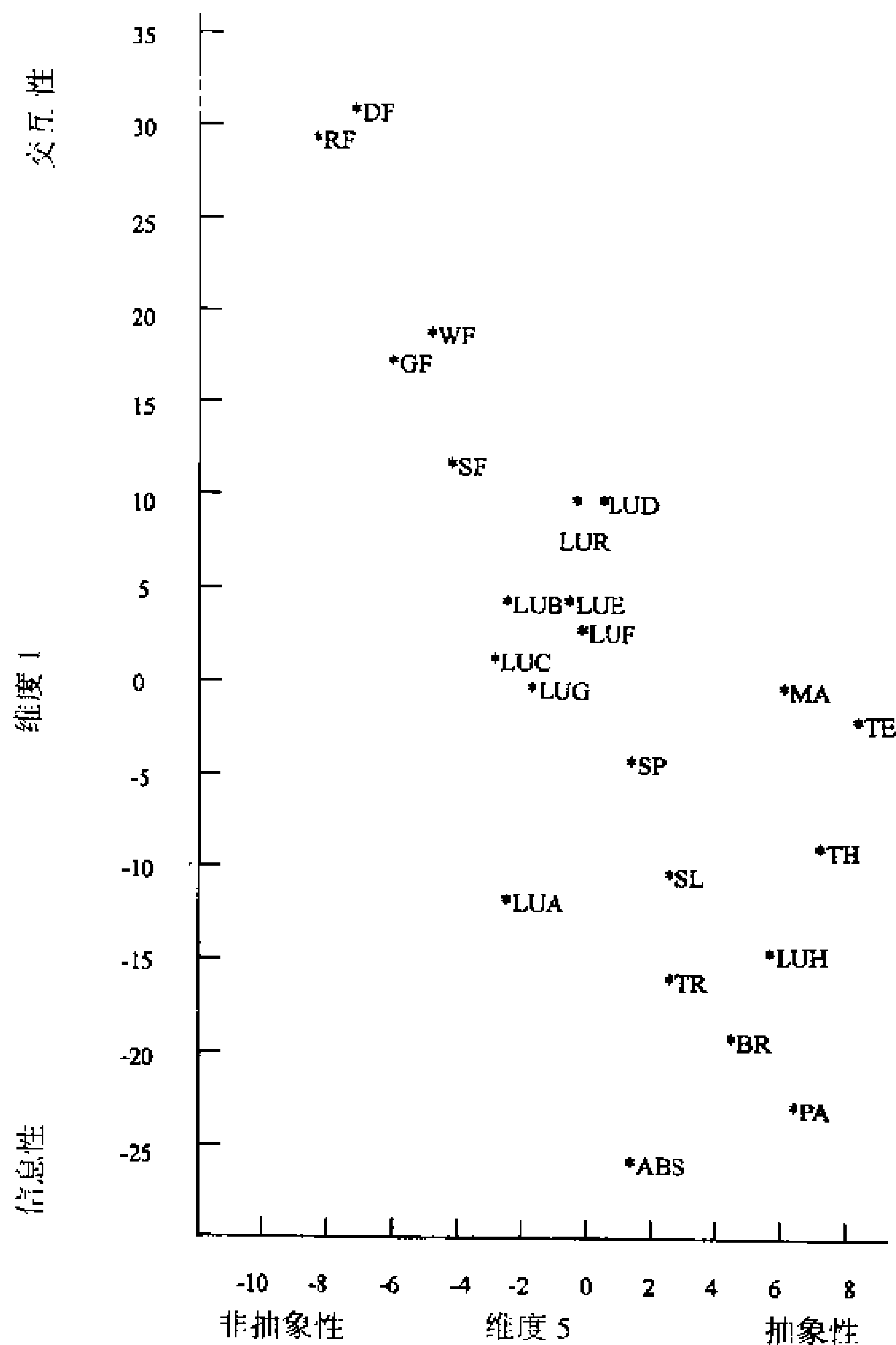
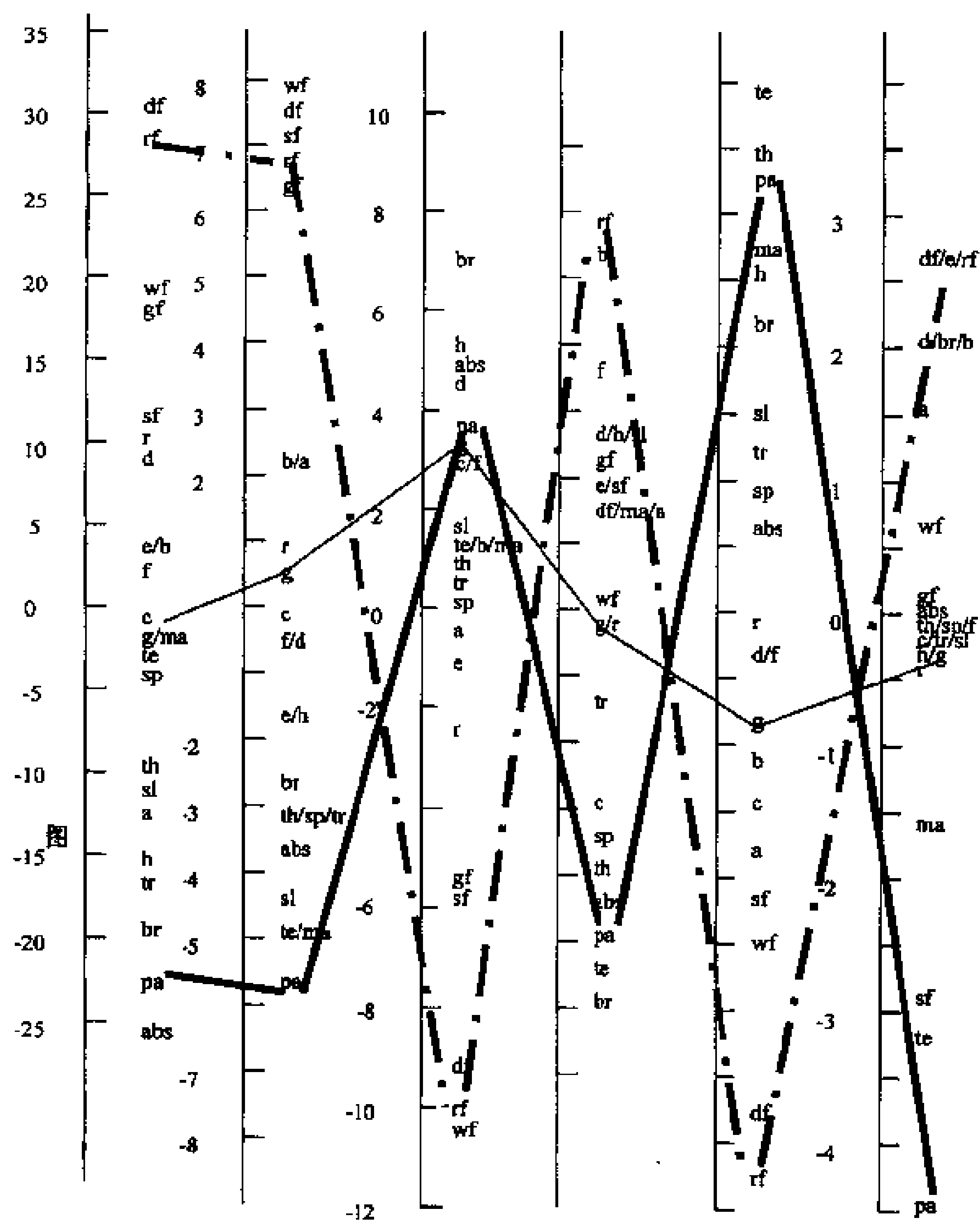


图 7.2 23 个文类在维度 1 及 5 上的分布

当六个维度都考虑进去时，则可对这些文类间的差异作出更全面的描述。图7.3为三个文类在六个维度上的分布情况。该图描述了每种文类的总体特征及任意两种文类的相互关系，并重点标出了三种文类：爱情小说、自传及专利的关系。每种文类在六个维度上都有不同的特征。爱情小说交互性强（D1）、有较强的叙述性（D2）、所指依赖场景（D3）、劝诱性强（D4）、给出的是非抽象性信息（D5）、有很强的即时信息详述目的（D6）。专利各维度上的特征恰好与爱情小说相反，与其形成了强烈的对照。自传的所指对场景有较强的依赖性、内容偏于非抽象性，在其他方面得分均居中等水平，特征不显著。依此类推，读者可对其他文类的特征及其相互关系进行描述。从23个文类在六个维度上的分布关系可以看出，学术英语在其中五个维度上，特别是维度1、维度2及维度5上，位置很突出。所有JDEST语料库语料在维度2（"叙述对非叙述"维度）上得分都低；在维度5（"抽象对非抽象信息"维度）上得分都高。在维度1（"交互性对信息生成"维度）上，除与来自LOB语料库的语料略有重叠外，其得分也大都较低。在其他维度上，尤其是维度3上，来自JDEST的语料与来自LOB的语料重叠的部分要多些。说明两者在这些维度上区别要小于另外三个维度。来自JDEST的九个文类位置不同，但总是很靠近，说明它们互相间存在语体风格的差异，同时又有共性。它们之间的差别小于它们与来自LOB语料库的语料之间的差异。

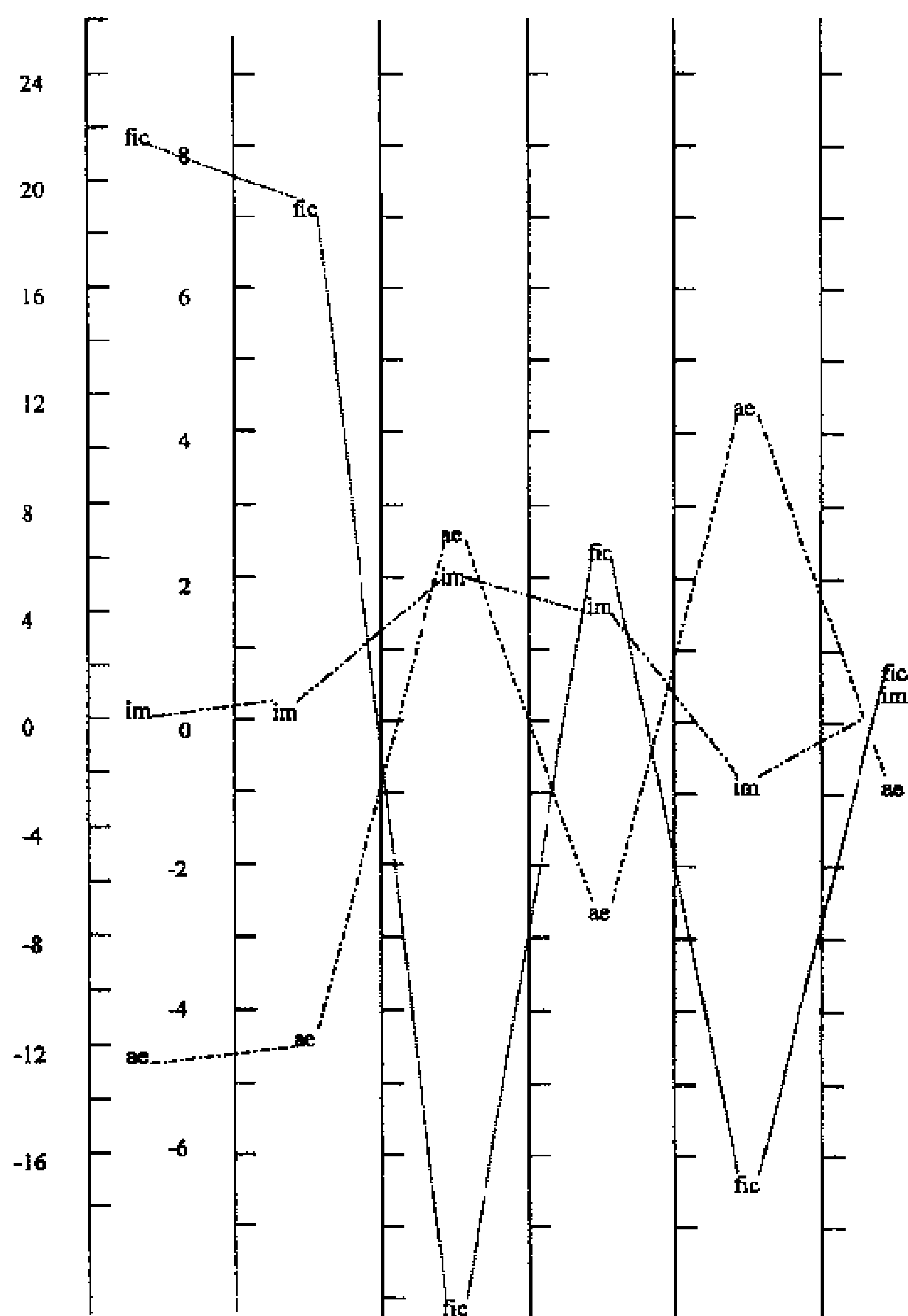
从图7.3可见，这23个文类显然分成三组，即来自JDEST的语料、来自LOB的语料中的小说部分和LOB语料中的其他部分。三者可分别定义为学术英语、小说及中间文类。小说在维度1、2、3和5上处在与学术英语完全相反的突出位置。但在其他维度上，学术英语和小说都与中间文类有较多的重叠。说明在这些维度上三者共性多些。

因此我们可以计算三组文类在六个维度上的平均得分（见表7.11），并将结果在各维度上标出。如图7.4所示。



注：符号所表示的文类见表 7.9，图中 a～h 分别为 LUA～LUH 的缩写，所标出的三种文类为 rf —— 爱情小说、g —— 自传、pa —— 专利。

图 7.3　三种文类在六个维度上的分布



注：ae —— 学术英语，fic —— 小说，im —— 中间文类

图 7.4　三组文类在六个维度上的平均分

表 7.11 3组文类在每个维度上的平均得分

维度	学术英语	小说	中间文类
1	-12.3	21.6	0.3
2	-4.1	7.2	0.1
3	2.6	-8.2	2.0
4	-2.8	2.3	1.6
5	4.4	-6.3	-0.9
6	-1	0.7	0.6

图 7.4 显示 3 组文类在维度 1、2、3 和 5 上有很大差别，而在其他维度上差别小一些。根据维度的定义，维度 1 包含代表交互与概括性内容的语言项目及那些代表细致与精确内容的语言项目。在该维度上，小说位置比其他文类都高，学术英语位置最低。维度 2 由标记叙述性和非叙述性目的的语言项目构成。小说在该维度上得分很高，说明其叙述性目的很强；而学术英语得分最低，表明其非叙述性目的很强。分数居中的文类叙述性、非叙述性特征都不显著。维度 3 定义的是所指的明确性程度。小说在所指方面对场景依赖性很强，而其他一般英语文类及学术英语在这方面的要求不显著，尤其是一些学术英语文类更强调所指的明确性。维度 5 有抽象性信息特征，如被动式和过去分词从句等。该维度同样将所调查的 23 个文类区分为 3 组，即学术英语、小说和中间文类。学术英语是最抽象的，小说抽象性最低，而中间文类的抽象性居中。维度 1、2 和 5 将 23 种文类明确地区分为 3 类，尤其在维度 2 上，3 组文类没有任何重叠，说明它们在叙述性方面有绝对的差异。在其余几个维度上，23 个文类间也有系统性的差异，不过相对小一点。

简言之，学术英语和小说在各维度上的位置很突出。小说有较强的交互性(D1)，较强的叙述性(D2)，所指依赖场景(D3)，提供非抽象信息(D5)。其 D4(劝诱性)和 D6(即时信息详述目的)上的特征虽不显著，但还是高于学术英语。学术英语在 D3、D4 和 D6 上

特征不太显著，其特点是信息生成的(D1)、非叙述性的(D2)和极其抽象的(D5)。中间文类位于各维度的中部，表明与另外两组相比，中间文类中所调查的67项语言项目的使用频率居中。

这些文类在图上的分布说明学术英语是有独特语体风格的。贝博所定义的六个维度能够从功能角度出发从整体上描述学术英语的语体特征。根据六个维度各自包含的语体标记及学术英语在各维度上的位置，很容易确定模型中涉及的语言项目在学术英语中的使用情况与在其他类英语书面语中使用情况的差异，从而建立起学术英语语体风格特征的大致轮廓。

恩克韦斯特(1964)把语体风格定义为语言项目的概率的集合。他认为，语体标记指仅出现在某些语境，或极多/极少出现在某些语境的语言项目。出现在同一文本中的语体标记构成该文本的语体标记集。相关但不同语境中出现的大量文本共有的语体标记构成一个主要的语体标记集。含有主要语体标记的文本属于同一主要语体。该思想与贝博的MD/MF模型的理论依据是不谋而合的。如果考虑到给定语言的所有语境，并分析足够多的文本，即可根据语体标记出现的特点定义该语言的主要变体及其相互关系。在确立重要语境类别、语体标记、语体标记集及主要语体的定义后，这些定义可在指导教学及更有代表性的语料库的建立等诸多方面发挥重要的作用。

第八章　学术英语中的语义韵研究

引　　言

语料库证据驱动的词语搭配研究表明，有些词项的搭配行为显示着一种特殊的趋向：它们习惯性地吸引某一类具有相同或相似语义特点的词项，与之构成搭配。由于这些具有相同或相似语义特点的词项习惯性地、反复地与关键词项在文本中共现，关键词项也就被"传染"上了有关的语义特点，它的整个语境内也就弥漫着一种特殊的语义氛围。这就是语义韵 (semantic prosody)(Louw 1993: 156- 159)。词语搭配形成的语义韵大体上可分为消极语义韵 (negative prosody)、中性语义韵 (neutral prosody) 和积极语义韵 (positive prosody) 三类(Stubbs 1996: 176)。在消极语义韵里，所研究的词项吸引的词语几乎都具有强烈或鲜明的消极语义特点，它们使整个语境弥漫着一种强烈的消极语义氛围。积极语义韵的情况则正好相反：所研究的词项吸引的几乎都是些具有积极语义特点的词项，由此形成一种积极语义氛围；在中性语义韵里，搭配词语的语义特点既不消极，也不积极。也可以说，既有一些消极涵义的词项，也有一些积极涵义的词项，呈现出一种错综的语义特点。绝大多数英语词的搭配行为呈现中性语义韵，而一些词项具有强烈的消极语义韵，另一些词项则有明显的积极语义韵。

本章将用词语搭配研究的一般方法，也就是说，将搭配跨距界定为 +4/-4，提取有关搭配词，同时再参照一定的类联接，对搭配词的语义特点观察和描述。文章将探讨学术英语文本中词项的两类语义韵，即消极语义韵和错综语义韵。前者包括 *cause*、*hap-*

pen、*incur*、*incidence*、*utterly* 等词的搭配行为。后者有 *probability of* 的搭配行为。这些都是专业文本中的次技术词。研究证据来自400万词次的上海交通大学JDEST语料库；该语料库已被学术界公认为学术英语语料库。另外，本研究将参照100万词次的英国英语LOB语料库和100万词次的美国英语BROWN语料库证据，进行对比。本章将依次讨论这些语义韵的形式以及语义韵的利用和语言学意义等问题。

1　消极语义韵及其有关词项

如果一个节点词的搭配词绝大多数都具有消极的语义特点，那么，该节点词就会具有一种消极语义韵。辛克莱(Sinclair)注意到，与 *set in* 这个词组动词搭配并做其主语的词大都是些消极涵义的词项。除了少数几个表示“天气”的词和 *reaction*、*trend* 等词之外，*set in* 的主要搭配词包括 *rot*、*decay*、*malaise*、*ill-will*、*decadence*、*impoverishment*、*infection*、*prejudice*、*vicious* (*circle*)、*rigor*、*mortis*、*numbness*、*bitterness*、*mannerism*、*anticlimax*、*anarchy*、*disillusion*、*disillusionment*、*slump* 等。所有这些词均指生活中不受人欢迎的事物或者某种令人不快的事情(1991:74－75)。洛悟(Louw 1993)和斯塔布斯(Stubbs 1996)分别研究了普通英语中有关词项的语义韵。笔者对JDEST语料库的词语搭配研究发现，专业文本中的一些词也具有明显的消极语义韵。

1.1　*cause* 一词的强烈消极语义韵

1.1.1　有关距位上的搭配词

cause 在JDEST语料库中共出现了949次。这949个实例表

明，节点词周围的词项大都具有消极涵义。为了便于研究，笔者从该语料库中随机抽取了 110 个实例进行观察。其中，有 5 例表示“事业”之义，如 *the cause of peace* 等。其他的 105 例分别表示两个相关的意义，“原因”之义和“致使某事发生”之义。由于本次研究关切的焦点是词形 *cause* 用于后两种意义的搭配行为，所以与之无关的五个实例将不予考虑。在剩下的 105 个实例中，*cause* 所用于的类联接有(1) V + N + to-V，如 *cause the machine to move*；(2) V + N、如 *cause errors*；(3) N + PREP + N，如 *cause of failure*、*cause for concern* 等。鉴于篇幅所限，现每隔两行取一行将其中的 35 个词语索引显示如下：

1	ourette's syndrome, which can	cause a chronic involuntary tic and affects
2	mall flow in each channel may	cause the element to switch unexpectedly. T
3	achine controller which would	cause a modified milling machine to move in
4	Then, errors in setup would	cause errors in surface generation. The gra
5	nnot be force-balanced and so	cause the inability to fully force-balance
6	ize). The Mekong Cascade will	cause environmental disruption across the w
7	able can rise sufficiently to	cause transient waterlogging. 5. Compared w
8	ficant variations in load may	cause severe speed fluctuations commonly re
9	f the second element and will	cause load sensitivity or even oscillation.
10	some symbolic format; it can	cause the controlled machine to execute the
11	ted by constant operation can	cause high friction and positioning errors.
12	ay feel for someone close may	cause intense guilt and self- torment if th
13	table and therefore likely to	cause greater disruption to productivity. M
14	ilot and counterbore, and can	cause breakage of either one. To avoid brea
15	free of undesirable odors or	cause irritation to the skin. They should a
16	times wide-scale food aid can	cause problems rather than solve them.
17	in taking notes, one that can	cause grief later on in writing, is to dist
18	shapes, but the vibration can	cause noise, wear and compatibility problem
19	rampant inflation and thereby	cause even greater damage to most of the ci
20	e shoulder strap itself could	cause neck injury in a side-on crash.
21	injury or surgery, can also	cause embarrassment by, for instance, the

22	ing practice and will usually	cause the forging to be rejected. The fourt
23	sening; a slipping wrench can	cause smashed fingers or other injuries to
24	ons. Air pollution injury can	cause early senescence or leaf drop. Stems
25	tude of organic chemicals may	cause taste and odor problems in water, wit
26	differential is sufficient to	cause an increase in the river temperature,
27	rges of these types certainly	cause the more spectacular effects on river
28	rst, toxic contaminants could	cause the total eradication of life within
29	it would have a bad odor and	cause hoarseness. The ancient Romans constr
30	dia. Small hard particles can	cause g uin hdcoyo89 permanent disc defects
31	high by local standards would	cause rampant inflation and thereby cause e
32	repeated cycles of strain can	cause fatigue damage. but instead of strain
33	lling machine where they will	cause scoring and rapid premature wear. Cle
34	c. Hydrogen chloride can also	cause severe corrosion problems in boilers.
35	e abrasive in nature and will	cause accelerated wear on pumps and sludge-

我们将以上述 35 行词语索引为基本依据，分析词形 *cause* 在各个跨距位置上的搭配行为。首先，在 N－1 位置上，也就是紧靠节点词左边的第一个距位上，相当多的搭配词都是情态助动词，如 *can*，*may*，*will*，*could*，*would* 等。其中，*can* 出现了 12 次，*will* 4 次，*may* 4 次，*would* 3 次，*could* 2 次。另外，还有 *also*（21、24）、*usually*（22）、*certainly*（27）、*to*（7、13、26）等。情态助动词一般在语篇中表达一些情态意义，如建议、请求、许可和告诫。但是，仔细观察会发现，那些除情态助动词之外的搭配词也在不同程度上实现着情态意义。如 21 行和 24 行词语索引中的 *also* 实际上是 *can also* 的一部分，其中情态助动词 *can* 出现在了 N－2 位置上，而 7、13、26 行词语索引中的 *to* 则分别是 *sufficiently to*、*likely to*、*sufficient to* 的一部分。它们实际上也是英语中的情态意义手段。其他搭配词的情况也是如此。所以，N－1 位置上的搭配词大体上可概括为两类：明显的情态动词和不太明显的情态助词，各搭配词出现的频数如图 8.1 所示：

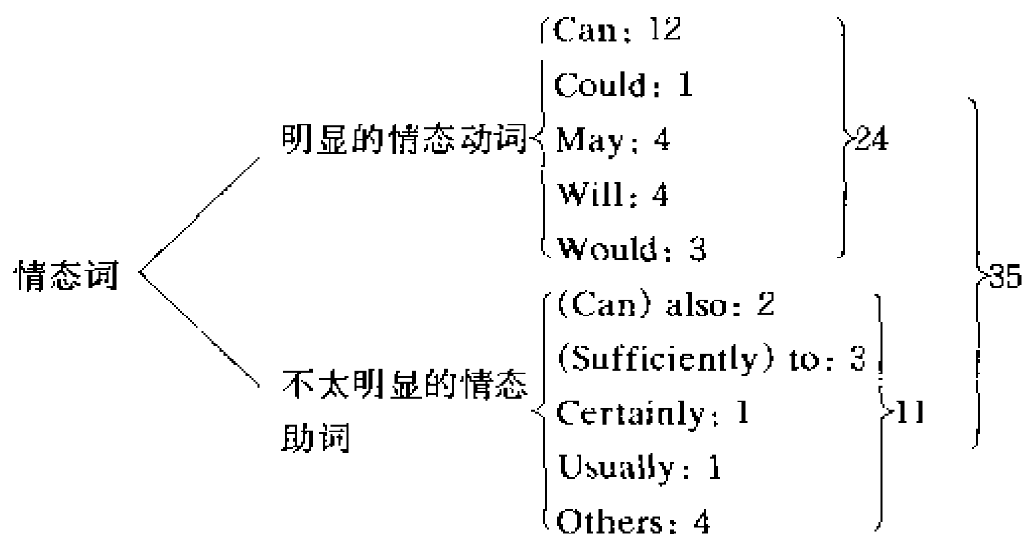


图 8.1　在 N-1 位置上的搭配词

如图 8.1 所示，在 N－1 位置上，词形 *cause* 有 65%(24/35)的搭配词都是情态助动词，35%(11/35)的搭配词是英语中表示情态意义的各种不同手段。我们可将这些搭配词的情况抽象为"表示情态意义的类联接"(a colligation of modality)。

也就是说，在 N－1 位置上，词形 *cause* 类联接于 (colligates with) 情态结构，其中 *can*、*may*、*will*、*would* 等助动词有 65% 的概率被语言使用者选择来实现这个类联接，其他各种情态手段有 35%的概率被选择。

下面分析 N－2 至 N－4 位置上搭配词的情况。由于 *cause* 在目前的实例中都是动词，且主要用于 V ＋ N 这个类联接中，所以在 N－2 至 N－4 位置上，我们有理由期待做其主语的名词词组大量出现。而且，我们可将 N－2 至 N－4 位置上的搭配词放在一起检查，而不是按照距位逐一检查。由词语索引可知，在 N－2 至 N－4 位置上，主要的名词词组有：*syndrome*、*errors*、*undesirable odors*、*the vibration*、*rampant inflation*、*injury or surgery*、*slipping wrench*、*air pollution injury*、*organic chemicals*、*toxic contaminants*、*a bad order*、*repeated cycles of strain*、*Hydrogen chloride*、*small hard particles* 等。

不难发现,这些名词词组都是各个科技专业领域里的技术词汇和次技术词汇,它们的一个共同特点就是"指称各个专业领域里令人不快的事物",或者说,指称那些不妨称之为"问题"的一些消极事物。没有一个词组具有积极涵义,全都是"问题类"的词汇。这就是N-2至N-4位置上搭配词的主要特点。另外一些搭配词似乎属于中性词语,如 *The Mekong Cascade* (6)、*variations* (8)、*second element* (9)、*these types* (27)等等。在目前的语境内,我们难以判断它们所指何物。然而,如果观察其更大一点的语境,就会发现下列详细信息:词语索引6的 *The Mekong Cascade* 指的是缅甸政府大规模从湄公河水系抽水,用以灌溉农田的计划;在生态保护主义者看来,这将是灾难性的事件。其他几个索引的详细语境如下:

1. Because leakage may approach or exceed the flow required to turn the motor, any significant *variations* in load may cause severe speed fluctuations commonly referred to as cogging. (8) ...
2. circulation around channels G and F has an adverse effect upon the stability of the *second element* and will cause load sensitivity or even oscillation. (9)
3. ... pollutants as being oxygen-depleting or directly toxic to the river fauna. Discharges *of these types* certainly cause the more spectacular effects on river systems. (27)

由这些详细语境可知,*variations* 与 *leakage* 密切相关;*second element* 并不是索引行9中 *cause* 一词的主语,主语是 *adverse effect*;*these types* 指的是*pollutants* 和*toxic*,是 *Discharges* 一词的修饰语。那么,这些表面上似乎中性的词汇实际上并不中性,而是以这样或那样的方式与一些消极词汇相联系,因此也具有

消极涵义。在 N－2 至 N－4 位置上，我们还可以看到其他一些异质搭配词，如 *be force-balanced* (5)、*some symbolic format* (10)、*by constant operation* (11)、*free of undesirable odors or* (15)、*it would have a bad odor* (29)等。这些都是不同的语法结构影响的结果。但总的来说，名词词组在 N－2 至 N－4 位置上占着主导地位。所有这些名词词组都有一个共同的语义特点，即“令人不快”或“不合意”。那么，我们可以有把握地说，从 N－2 至 N－4 位置上的搭配词都具有“令人不快”或“不合意”这一语义趋向(a semantic preference of undesirability)。

现在，我们来检查词形 *cause* 的右搭配词的一般情况。右搭配词的情况要比左搭配词有规律得多，因为 *cause* 主要用于两个类联接：V + N + to－V 和 V + N。我们不妨将这两个类联接内的搭配称为核心搭配(core collocation)，参照这两个类联接，提取所有核心搭配的搭配词，而不必一个位置接着一个位置地去分析。V + N + to-V 类联接有两个实例：*cause the element to switch unexpectedly* (2)和 *cause the forging to be rejected* (22)。这两个实例明显地表达专业领域里应当避免的“事故”或“问题”类的事物。其余的 33 个实例全部属于 V + N 类联接的具体实现。该类联接内名词搭配词的特点也显而易见。没有一个词项具有积极涵义；所有的名词搭配词都指专业领域里消极的“坏事”，如 *chronic involuntary tic*、*environmental disruption*、*rampant inflation*、*disruption to productivity*、*total eradication of life*、*permanent disc defects*、*irritation to the shin* 等。V + N 类联接内的主要名词搭配词可用图 8.2 表示：

1.1.2 *cause* 的语义韵结构及其启示

至此，情况已十分明显，V + N + to-V 和 V + N 两个类联接内的搭配词都是些消极涵义词汇，指那些在专业领域里应当避免的“问题”或“事故”或可能的消极后果。这些特点与辛克莱讨论的

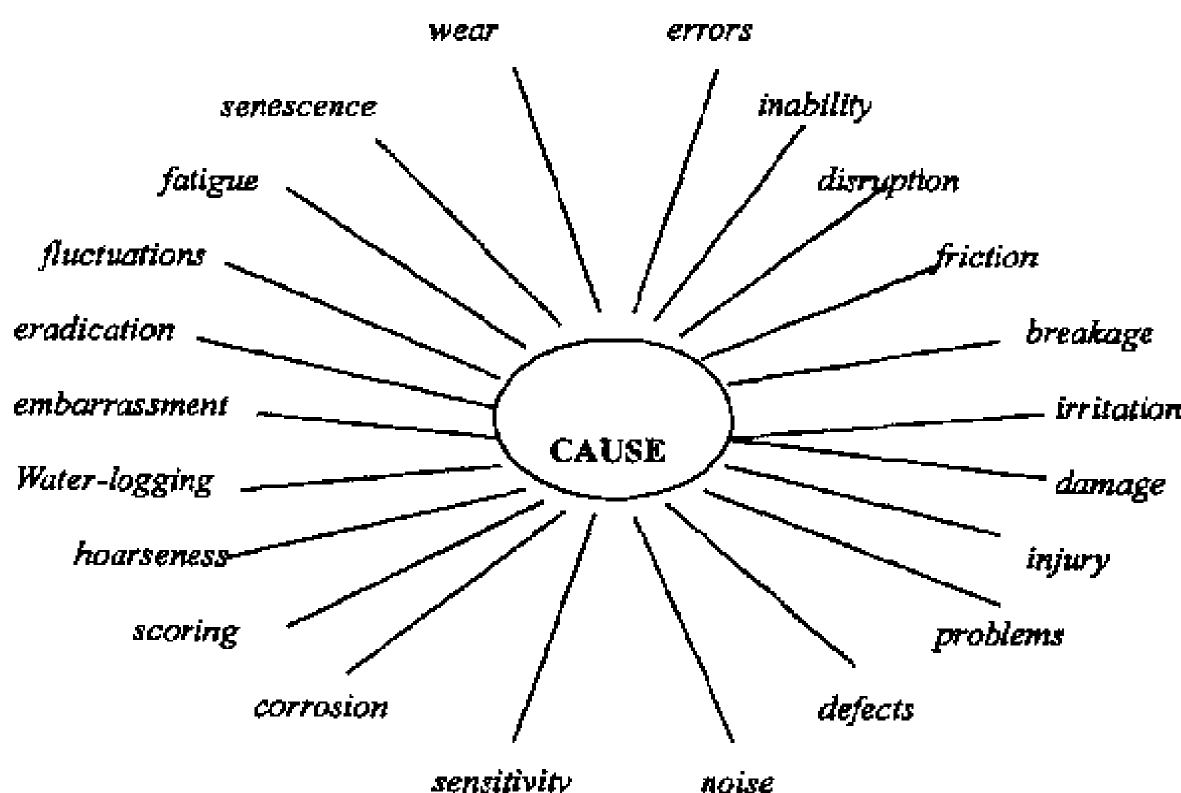


图 8.2 *cause* 一词的右消极涵义搭配词

set in 的主语搭配词的情况十分相似。斯塔布斯(1996) 也注意到，*cause* 在普通英语文本里的主要搭配词有 *accident*、*concern*、*damage*、*death*、*trouble* 等。这些词与我们研究的专业文本里的上述搭配词的语义特点一致，只是在专业文本里，核心搭配里的词项都是些技术词汇与次技术词汇。那么，现在回过头来看 N－1 位置上的情态类联接，其意义就较为容易解释了。该情态类联接必定与"避免"核心搭配里的"事故"与"后果"有意义上的密切联系。也就是说，这个情态类联接是用来发出一种"告诫"，提醒人们注意或避免有关的"事故"或"后果"。那么，整个语境的语义韵也就是实现这一"告诫"意义。如下面的实例所示：

6. The Mekong Cascade will cause environmental disruption across the whole peninsula.

8. Signficant variations in load may cause severe fluctuations.
25. Multitude of organic chemicals may cause taste and odor problems.
24. Air pollution can cause early senescence or leaf drop.

将上述所有分析综合一起，*cause* 的语义韵则是"告诫的语义韵"，其结构由"表示'不合意'的语义趋向"加"情态类联接"加"核心搭配"构成。如图 8.3 所示：

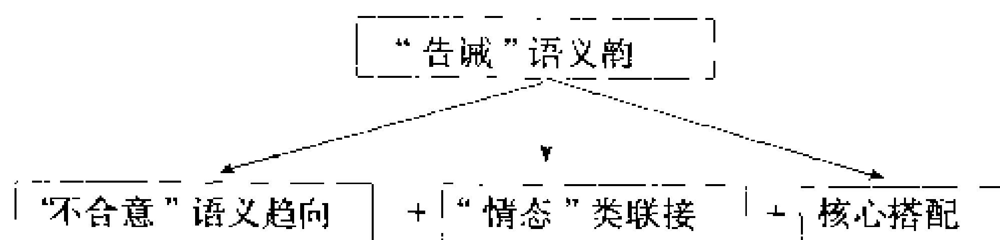


图 8.3　*cause* 一词的语义韵结构

词形 *cause* 语义韵结构的建立具有极大的语言学启示。它表明，节点词在限定语境里的所有搭配词的选择都受制于语义韵。语义韵制约着语义选择趋向，语义选择趋向控制着 N－2 至 N－4 位置上所有的词语选择，语义韵也制约着 N－1 位置上类联接的选择，而核心搭配里的词项选择也服从于整个语义韵的意义。这表明，语言交际中的形式选择，无论是句法结构还是词语项目，都服从于意义的实现：意义第一，形式第二。而传统的语言学描述体系却总试图使人们相信"句法第一，词汇第二"；词汇仅是用来填充一定语法结构的材料，句法是居于统治地位的制约形式选择的规则。语义韵研究的发现无疑对这种根深蒂固的观念提出了挑战。语义韵的发现还揭示了"搭配和谐"这一重要的原则和机制。一旦词项发展起了自己的语义韵，一定的语义氛围也就随之产生。位于节点词语境里的各词项都必须符合语义韵的要求，与整个语义氛围和谐。只有那

些具有一定语义特点的词项才可能被吸引到语境里与节点词共现，与语义氛围冲突的词将被排斥在语境之外，除非是为了创造一定的修辞效果。搭配和谐的力量如此之强，以至于整个语境里所有位置上都弥漫着特定的语义氛围，甚至限定跨距之外的一些词也或多或少地受到语义氛围的影响。由此，不仅核心搭配里的词项属于同一的语义范畴，那些与节点词关系不太紧密的搭配词也受到语义上的传染。正如辛克莱所言，"语义韵将词项与其环境结合一体，起着主要的作用。它几乎表达了词项的功能——表明词项的其余部分如何从功能的角度解释。没有语义韵，一串词仅有一种"意指"，而不是用于可行的交际之中"("... the semantic prosody has a loading role to play in the integration of an item with its surroundings. It expresses something close to the function of the item — it shows how the rest of the item is to be interpreted functionally. Without it, the string of words just "means" — it is not put to use in a viable communication")(1996:87)。

证据似乎表明，词项的意义就在语境之中，存在于与之结伴的别的词项之中。节点词的意义与语境中搭配词的意义交织一起，难解难分。

1.1.3 N + PREP + N 类联接中的搭配词

上面讨论的是 V + N + to-V 与 V + N 两个类联接中搭配词的情况。稍稍观察一下 N + PREP + N 这一类联接中的搭配词，我们便会发现，*cause* 在这里的消极语义韵也极为明显，如下面词语索引所示：

1	and thermal stresses are the	cause of the appearance of machining stress
2	ved will help to identify the	cause of reduction in yield Laboratory stu
3	The second, and less obvious	cause of disruption, is the manipulation of
4	te oxidation reduction is the	cause of exfoliation which has been a serio
5	h information to discover the	cause of this intermittent oscillation at t

6	a 'bottleneck' machine is the	cause of batch and perhaps machine idle tim
7	ty of the mounting. The exact	cause of failure may vary from too much ove
8	temperature and the only real	cause of dissatisfaction with such high wor
9	strain which is an additional	cause of inaccuracy. These difficulties hav
10	e. That eventuality must be a	cause for anxiety. The alternative would be
11	now was searching to find the	cause of a cholera outbreak in the St. Jame
12	red to remove or suppress the	cause of the impact; this may involve such a
13	etal fatigue will be the only	cause of failure. Since this implies many
14	machine tools. The immediate	cause of failure when chatter exists is tha
15	comes, as it did in Eden, the	cause of our fall. The naturist novel retur

显而易见,搭配和谐的机制在起作用。正如前面两个类联接中搭配词的情况一样,N + PREP + N 中的名词搭配词全都是些消极涵义的词项:*machining stress*、*reduction in yield*、*disruption*、*exfoliation*、*intermittent oscillation*、*idle time*、*failure*、*dissatisfaction*、*inaccuracy anxiety*、*cholera outbreak* 等等。这些词项均指各专业领域里的"坏事"、"问题"、"事故"或人所不欲的"后果"。

1.2 其他有关词项的消极语义韵

在上面几节里,我们讨论了词形 *cause* 的消极语义韵;建立了语义韵结构并对限定跨距里的搭配词作了详细分析。其目的在于说明语义韵的功能及搭配和谐的重要原则。但这并不意味着,语义韵的研究都要采用同样的方法和步骤。这样做很费力气。一般情况下,较为经济和简便的办法是从界定跨距或类联接出发,观察节点词在跨距或类联接内的搭配行为,概括其可能的语义韵。笔者发现,专业文本里相当一部分词项,如 *happen*、*incur*、*incidence*、*utterly* 等,都有一定的消极语义韵。这里予以扼要讨论。

1.2.1 *happen*

happen 一词具有强烈的消极语义韵。在 JDEST 语料库中,

happen 一共出现了 101 次，常作 *happen* 主语的有名词、代词和 wh-疑问词。其主要类联接如下：

N + (AUX. +) V: Accidents happen. Things can happen.
PRON + (AUX. +) V: This may happen.
Wh-word + (AUX. +) V: What happens
Indefinite PRON + (AUX. +) V: Anything can happen.

名词主语都是表示"事故"、"灾难"、"不幸"、"问题"一类的词，常见的有 *accident*、*disaster*、*failure*、*misfortune*、*problem*、*unfortunate things*、*suicide* 等。作主语的代词主要是 *it*、*this*、*that* 等。不定代词作主语的情况有 *something*、*anything* 等。wh-疑问词作主语的实例主要是 *what happens*、*whatever may happen*、*What (will) would happen if* ... 等。如下列索引所示：

1	is taken to a logic 0 two things will	happen: 1. The operation of the flip-flop is i
2	y may argue, "who can tell what will	happen 20 years hence?" T3082F03A1988TE20 1.
3	control is simply assuring that things	happen according to plan. Surprisingly, this
4	countries and have considered what may	happen after 1992 as economic integration adva
5	attack, and apprehension that it could	happen all over again. Crown Prince and
6	ut just talking about it won't make it	happen. Along with the AFS organization, we,
7	the speculation. This is beginning to	happen; although relatively few of those who r
8	as all right as he knew it was going to	happen and there was no way to stop it. He
9	t everything was new, anything might	happen, and all life was before us. We look
10	d philosophical approach is the "let it	happen" approach, or the approach in whi
11	nevitable failure, but to let failures	happen as they may, and to continue t
12	andfather died — but think what would	happen as he read the obituary lists during th
13	ness. A commonplace incident that could	happen at any time, and which nonetheless cont
14	thic to euktic, a prayer that something	happen. Before the Fall their prayers were ap
15	for petroleum grades. We knew it had to	happen, but didn't feel it would be this soon

16	today's children. This eventually will	happen. but, for the present, what appears
17	dle nose. fig 9-5; if this does not	happen, do not continue but inform your superv
18	ecreased in another region. This may	happen. for example. if control rods are
19	e. If the reader knows what is going to	happen from the beginning, he or she immediate
20	entally, than the desire to make things	happen? I present to you History, the fabricat
21	tend to dilate. We wondered what would	happen if subjects with high DIT scores ind
22	om each other, The same thing would	happen if two south poles were brought near e
23	r can now ask the computer,"What would	happen if I ...?" and he will get a helpful a
24	what happens and also what ought to	happen. If the politics which emerge are not
25	ould have changed, although it could	happen. If one searches terms there is a

代词和 wh-疑问词作主语时，由于语境所限，我们很难确定它们的语义特点。然而，检查一下扩展语境，就会发现这些代词和 wh-疑问词均指一些“坏事”：

1. The second philosophical approach is the "let it happen" approach. The idea here is not to postpone the inevitable failure, but to let failures happen as they may.
2. ... individual adaptation acts as lubricants and thereby generate even deeper crises, although only much later will she elaborate exactly how this will happen.
3. It may happen that the data produces such long strings of consecutive zeros or ones that the receiver looses synchronization.
4. Great curiosity and desire to experiment often got him into trouble. When he was six, he set fire to his father's barn "to see what would happen". The barn burned down.
5. He would no doubt feel guilty when his grandfather died, but think what would happen as he read the obit-

uary lists ...

作主语的不定代词一般也指不好的事，或者与不好的事有关。如：

After the Fall, their prayers were apathetic, prayers that something not happen, that no evil appear in the night ... (JDEST Data)

上述引言取自关于人类始祖 Adams 和 Eve 堕落的一篇文章。引言显示，*something not happen* 指的是 *no evil appear*；*happen* 和语境中的 *Fall*、*evil* 等词有着一定联系。上述证据表明，*happen* 并非某些英语教材所讲的是 *occur* 和 *take place* 的同义词。在英语中，凡是 *happen* 的大都是些坏事，而 *occur* 与 *take place* 则无此语义韵。JDEST 和 BROWN 的证据都显示，*occur* 的主语既可以是具有消极涵义的词，也可以是具有积极涵义的词；*take place* 的主语则是一些中性涵义的词，如 *event*、*incident*、*meeting*、*adventure*、*transaction*、*developments*、*investigation*、*discussion*、*transformation*、*ceremony* 等。

1.2.2 *incur*

COBUILD 词典将 *incur* 的意义解释为 *if you incur something unpleasant as a loss, the way in which you act or behave causes it to happen*.（Sinclair, 1987: 739）。其中，*unpleasant*、*cause it to happen* 等词语揭示了该词的消极语义韵。下面是从 JDEST 语料库中提取的 *incur* 和 *incurred* 两个词形的部分词语索引：

1　ras the state is only allowed to　incur debts in a period of transition. In a just
2　r day for a commuting trip does not　incur extra depreciation costs, since the car is

3	ng that you will not feel guilty or	incur other nonmonetary costs from not putting mon
4	— without imposed controls — and	incur no administrative expenses. ~T3229K90A1988P
5	to buy land the state must however	incur debts. This gives rise to a feasibility p
6	he parasite's energy budget and may	incur the risk of dislodgement in an environment w
7	nsive dispersants, which in turn	incur additional heating costs because they have t
8	drugs is without toxicity, and all	incur considerable expense. Selection of a specifi
9	information bit stream so that some	incurred errors at the receiving end can now be re
10	to determine the extent of error	incurred by assigning average values to certain
11	me amounts. To eliminate lost time	incurred while prototype castings are formed, in
12	nected with severe weight penalties	incurred by the need to shield and cool parts f th
13	the risk of litigation for injuries	incurred during work or for aggravated injury is e
14	ower total investment than might be	incurred through the acquisition of similar data p
15	Operating expenses are those costs	incurred in regularly using the equipment. In acc
16	n"which states that losses are	incurred when a response or"quality characte
17	To do this,he classified the waste	incurred in the production process into the fol
18	to ascertain energy usage and costs	incurred in the heat treatment of iron castings.
19	nvestment is the sum f those costs	incurred until the equipment is ready to produce o
20	associated professional schools who	incurred a work-related injury or illness during t
21	only limited transformation will be	incurred by the reactor vessel material in the
22	easonable profit and damages. Costs	incurred by the owner must be certified by the ar
23	rofits and losses which operators	incurred during these two years. The study conclu
24	evel of costs which will be	incurred and the revenues required by owners for n
25	eclined and substantial losses were	incurred owing to the high written down costs of

由上述词语索引可知,在 V + N 类联接内,常用的名词搭配词有:*debts*、*costs*、*expenses*、*risk*、*errors*、*losses*、*injury*、*loss*、*damage*、*burden*、*displeasure*、*lost time*、*penalties*、*waste* 等。其中,*incur debts*、*incur costs*、*incur losses* 频率极高,可以视为半固定词组。*incurred* 还可出现于 ADJ + N 类联接内,如 *the incurred costs*, *the incurred errors*.

1.2.3 *incidence of*

词形 *incidence* 在 JDEST 语料库中出现了 206 次。经分析,该词形主要用于两个义项。第一个义项是"入射、入射角"(the way

a ray of light strikes a surface),是物理学上的一个术语。用于该义项时,常见的搭配有 *the angle of incidence*、*beam incidence*、*normal incidence*。不言而喻,这些词组的语义韵是中性的。第二个义项是"发生、发生率"(the extent to which something happens, or the number of times something happens)。用于第二个义项时,"incidence"主要用于 N - of + N 和 N + of + V-ing 两种类联接内。两个类联接的搭配实例共有 123 个。在 123 例中,只有 *connection gas*、*local recurrence*(2 例)、*rising tune*、*sales taxes*、*skills*(2 例), *the operating light beam*、*this event* 等名词搭配词和 finding, 共 10 例没有明显的消极语义特点,占全部实例的 8%;其他 92%的搭配词都明显地具有消极的语义特点。

下面,我们分析一下在几个不同领域里,N + of + N 类联接内 *incidence* 的名词搭配词的特点:

1) 医学领域。在医学专业文本中,*incidence* 的名词搭配词主要是表示"疾病"、"死亡"、"伤害"、"感染"和"反应"等意义的词项:如:*heart disease*、*lung cancer*、*stroke*、*blood flukes*、*bronchitis*、*typhoid*、*leukaemia*、*tumours*、*measles*、*T. B.*、*dental caries*、*pancreatitis*、*death*、*injury*、*ruptured E P*、*infection*、*skin reactions*、*reaction*、*nasty skin eruptions*、*HIV* 等。

2) 机械和造船领域。在这一类文本中,*incidence* 的名词搭配词主要是表示"机械故障"、"误差"、"缺陷"等意义的词项,如:*acoustic wave*、*deck wetting*、*welding defects*、*errors*、*defects*、*holes*、*lead poisoning*、*gun flashover*、*thermal fatigue* 等。

3) 一般的专业领域。在相当多的专业领域,incidence 的名词搭配词主要是表示"异常"、"事故"、"紧急情况"、"缺陷"、"不规则"、"崩裂"等意义的词项,如 *abnormalities*、*collapses*、*disruption*、*accident*、*error*、*defects*、*irregularities*、*risk factors*、*urgent requirements* 等。

4) 日常生活领域。在日常生活题材的文本中,*incidence* 的名

词搭配词主要是表示“失败”、“犯罪”、“死亡”、“仇恨”、“问题”、“堕落”等意义的词项：*failure*、*death*、*crime*、*theft*、*child abuse*、*malignancies*、*bullying*、*problem*、*side effects*；*nightmares*、*degeneration*、*trauma* 等。

在 N + of + V-ing 类联接内，主要的 V-ing 搭配词有，*cracking*、*poisoning*、*damaging*、*causing damage*、*blooding*、*casting surface defects* 等。

COBUILD 词典(Sinclair 1987:735)对 *incidence* 解释为 *the incidence of something is the extent to which it happens or the frequency with which it happens*。它没有提及该词的消极语义韵，但给出了一个典型的例句：*There is a high incidence of heart disease among middle-aged men*。从 JDEST 语料库的证据来看，*incidence* 几乎全是用来指各个专业领域里“令人不快”之事的发生率和发生程度，“好事”的发生率是不用 *incidence* 来表示的。这就是其典型的搭配行为。因此，我们可以有把握地确立该词的消极语义韵。限于篇幅，下面只提供 20 行词语索引，供读者参照：

1	there was a steady decrease in the	incidence of tuberculosis, but between 1990 a
2	of transmission resulting in zero	incidence of the disease, and the total disap
3	of which have been related to the	incidence of the disease 35,36 and confoundin
4	ation. On the other hand, the high	incidence of local recurrence in stage I obse
5	probably contributes to the higher	incidence of bile duct carcinoma. CLINICAL
6	are not economical because the low	incidence of flashovers does not justify the
7	om Latvia of a rising trend in the	incidence of this pathogen. HIV and AIDS By t
8	It is also shown to influence the	incidence of cracking and micro-fissuring in
9	ypass graft implantation, and the	incidence of myocardial infarction and death
10	anised, and this would reduce the	incidence of disruption. Physical parts contr
11	ely applied "genetic" tests. The	incidence of the risk factor has not changed,
12	ubmerged arc welds also have a low	incidence of defects but the fatigue strength
13	s and plumbing resulted in a high	incidence of lead poisoning among t he upper
14	Aspirin reduces significantly the	incidence of stroke in patients suffering tr

15 ld be reassured, however, that the incidence of malignancy is extremely low. Tr
16 ten associated with cirrhosis. The incidence of such bleeding ranges from about
17 . Particular note was taken of the incidence of common welding defects such as s
18 h pernicious anemia have a higher incidence of gastric carcinoma, this associat
19 y. Other questions include: is the incidence of disease uniform throughout the c
20 al cancer, and observations on the incidence of colon cancer in some population

1.2.4 *utterly*

utterly 意为 *totally* 或 *completely*。在 JDEST 语料库中,该词主要用于两个类联接内,ADV + ADJ 和 ADV + V。一般来说,*utterly* 之后的搭配词可以是中性语义的词项,如 *spontaneous*、*aware*、*different*、*absent*、*impossible* 等。但其典型的搭配行为却显示较强的消极语义韵。

在 ADV + ADJ 类联接中,典型的形容词搭配词有:*impossible*、*different*、*extraordinary*、*astonished*、*mistaken*、*oblivious*、*selfish*、*serious*、*unconventional*、*naive*、*miserable*、*confusing*、*absurd*、*compelling*、*helpless*、*noisome*、*inexcusable*、*outrageous*、*stupid*、*powerless*、*detached*、*deliberate*、*dumb*、*exhausted*、*obedient*、*unconvinced*、*unlike*、*unfamiliar*、*disgraceful*、*incompetent*、*inadequate*、*incoherent*、*irrational*、*useless* 等。在 ADV + V 类联接中,典型的动词搭配词有:*withhold*, *brutalized*, *defeated*, *deflated*, *rejects*, *forgot*, *abandon*, *disgraced* 等。如下面的索引句所示:

1. He grew up in a hostile world that utterly defeated him.
2. Such a complete regenerative cycle is utterly impossible.
3. These people were utterly mistaken as to the real bearing of the social problem.

4. He was able to make the most outrageous jokes while appearing utterly serious.
5. His expression was cold, utterly unbearable.
6. He was mad and she felt utterly weary.
7. Homoeopathy utterly rejected the weapons commonly used against disease.
8. They were all utterly confounded and then became embarrassed.

2 混合语义韵：*probability of* 的语义韵分析

混合语义韵(mixed prosody)指的是这样一种词语搭配现象，即节点词的搭配词既有积极涵义的词项又有消极涵义的词项，整个语义轮廓呈现错综复杂的情况。混合语义韵和斯塔布斯说的中性语义韵大致相似，但前者强调两类截然不同的语义特点，不一定有中性的语义特点。JDEST 语料库的证据显示，相当多的次技术词具有混合语义韵。*probability* 就是其中之一。该词在 JDEST 中总共出了 277 次，其中 115 次是用于 N + of + N 和 N + of + V-ing 两个类联接中。现将其中的 30 行词语索引显示如下：

1	ertainty. will survive, with the	probability of 0.99 or whatever, over t
2	ning two elements, each having a	probability of 1/2. For a six-digit sequence, w
3	nt yield decompose with a	probability of 10^(-3) to 10^(-4) or greater pe
4	, such that there is only a	probability of 5 percent that the strength
5	ined? The authors stressed the	probability of a decline in the growth rate a
6	devoted to insurrection and the	probability of a successful insurrection, and t
7	oyees, fewer issues and a lower	probability of a pool of inhouse computer skil
8	s function LA (a), which is the	probability of a leak developing in a tank of a
9	an extremely small but non-zero	probability of a highly magnified boom f

10	maged. T3044F A1991PE40 The	probability of a successful insurrection depend
11	nd conducted fracture tests, the	probability of a complete, catastrophic failure
12	nd because it would increase the	probability of a successful insurrection. In a
13	devoted to insurrection and the	probability of a successful insurrection are p
14	. Customer security. The	probability of a capacity deficiency is the
15	is considered there is a finite	probability of a reduction in the ion's t
16	Considering only the	probability of a deficiency ignores the fact
17	1) the case where the transition	probability of a change in internal energy upon
18	ılable regarding the statistical	probability of a 100-yr flood to allow for accu
19	is also shown in Fig. 13.22. The	probability of a 1-cm or larger object strikin
20	diving condition. Moreover, the	probability of a diver hyperventilation underwa
21	has been developed to assess the	probability of a patient's relapsing after prim
22	rimination curve would imply the	probability of a restricted range of colours. T
23	uum lines resulted in a	probability of about one gun in four surviving
24	ith the degradation in the	probability of accomplishing the mission. With
25	ase of the CMLs. CGL has a high	probability of acute transformation which is ly
26	if the worker underestimates the	probability of an increase in nominal wage infl
27	d in equal proportions (i.e. the	probability of anyone child belonging to either
28	peration is because of the lower	probability of avoiding blood transfusion durin
29	new generating plants, and the	probability of blackouts increases. The only de
30	Fig. 8 for CBN ground. For a 50%	probability of break a fatigue strength of 628

如上述词语索引所示，*probability of* 的名词搭配词大致可分为三类。第一类是中性涵义的词项，如数字：*0.99*（*a probability of 0.99*）、*5 percent*（*a probability of 5 percent*），以及 *change*，*occurrence*，*prototype structure* 等词语。我们不妨姑且将其称为 *occurrence* 类的词。它们的实例有 26 个，占全部实例的 23%。

第二类搭配词是些具有明显积极涵义的词项，如 *success*、*survival*、*accomplishing the mission*、*acquisition*、*winning* 等。这些"积极涵义词项"不妨姑且称为 *success* 类词项。它们总共有 11 个，占总搭配词的 9%。

第三类是具有明显消极语义特点的搭配词，指科技和专业人员不欲看到的“事故”、“故障”、“失败”、“问题”、“误差”、“混乱”等。我们现将这类“消极词项”放在 N + of + N 和 N + of + V-ing 两个类联接内讨论。

在 N + of + N 中，常见的名词搭配词有：*failure*（*the probability of failure*）、*accidents*、*error*、*loss*、*malfunction*、*exposure*、*deficiency*、*severity*、*rupture*、*decline*、*insurrection*、*reduction*、*blackouts*、*break*、*damage*、*rejection*、*cracks*、*unconsciousness*、*radiation*、*tails*、*impact*、*idles*、*interference*、*corrosion*、*disparity*、*relapsing default* 等。

在 N + of + V-ing 中，常见的搭配词组有：*causing an error*（*the probability of causing an error*）、*exceeding the heating effect*、*getting a ticket*、*producing the errors*、*rejecting*、*not detecting the failure*、*having a deficiency*、*making an error*、*slamming*、*unsettling chance events* 等。两类消极搭配词语共有 78 个，占 68%。在消极搭配词中，*failure* 和 *error* 频数最多，分别出现了 7 次和 5 次，姑且将这类词称为 *failure* 类的词。那么，*probability of* 搭配词的语义轮廓可概括如下：

probability of
- ① 中性涵义搭配词：*occurrence* 类 23%
- ② 积极涵义搭配词：*success* 类 9%
- ③ 消极涵义搭配词：*failure* 类 68%

这就是 *probability of* 搭配行为中的混合语义韵的情况。说它具有混合语义韵，是由于其搭配词既有一般的中性搭配词，又有明显积极涵义的词项，还有许多具有强烈消极涵义的词项，语义韵的总体呈现着较为复杂的情况。数据表明，语言使用者完全可以用该词来谈论任何事物或事件的概率，无论是好事还是坏事，成功还是失败。但是，数据也向我们揭示了一个重要信息。*probability* 具有一种极强烈的趋势和消极涵义词语搭配。当科技工作者在谈

论事件的概率时，他们有68%的概率来谈论坏事情发生的概率：事故、故障、问题、失败和差错的概率。相比之下，该词只有9%的概率被用来谈论好事情发生的概率：成功、完成任务的概率。

probability 一词的行为为什么会有这样一种“消极趋向”呢？或许，这与语言使用中人们的心理特点有关。一般情况下，人们总是将事情的好的一面视为当然，而注意的焦点则是事物的消极因素或成分。以乘飞机旅行为例。安全舒适的旅行是视为当然的，而人们则很可能会潜意识地考虑到飞机失事的概率。同样，在科技工作中，一切科技活动都旨在成功，旨在取得预定的效果和完成预定的任务。这些都是视为当然的事。但，科技人员却趋于考虑、预估和谈论失败的概率、发生事故的概率、出现问题和差错的概率。因而也就会采取措施来减少这样的概率，最大限度地争取成功。而 *probability* 在 V + N 类联接中的动词搭配词则反映了他们的这种努力，即对可能发生的事故概率进行估计、预测和强调，并降低、减少或确保某个事件的概率。如下所示：

V + N 类联接

estimate assess underestimate stress increase reduce lower decrease	} the / a **probability of failure**, etc.

而 DET + ADJ + N 类联接中的形容词搭配词则反映了科研人员对可能发生事故的概率所作的一般估计、判断或描述：平均概率、高概率、低概率、有限概率等。

DET + ADJ + N 类联接

a/n (the) { average / high / higher / great / greater / finite / low / lower / 50% / statistical / lowest } **probability of failure** (causing an error, etc.)

3 异常搭配与语义韵利用

语义韵是一种特殊的词语搭配现象。由于语义韵的作用,只有符合一定语义特点的词项才可能和某个特定的节点词相互吸引,共现于同一语境。任何异质词在语境中的出现,都可能产生异常搭配(unusual collocations),造成一种语义韵的冲突(prosodic clash),并破坏搭配和谐(collocational harmony)。那么,在一般情况下,本族语者总会选用典型的词语搭配来实现意义。这是词语的典型搭配行为之表现,属于语言使用的因循性(Hanks 1988)的一面。然而,语言使用又不仅仅是在刻板地遵约行事;它还有其创造性和灵活性的一面。在特定的语言交际环境中,出于某种特殊的目的,为达到某种语言交际效果,人们会故意蔑视规约,违反搭配和谐的原则,选用一些语义上冲突的词项来构成不同寻常的搭配或异常搭配,造成语义韵的冲突。这就是对语义韵的利用(the exploitation of semantic prosody)。

语义韵的利用和格莱斯(Grice)(见 Levision 1983：103)讲的说话者故意蔑视会话合作原则(the flouting of cooperative principles),以期产生会话含义(conversational implicature)的道理相同。按照格莱斯的理论,会话含义的产生主要是由于人们在字面上蔑视数量准则(Quantity maxim)、质量准则(Quality maxim)、关联准则(Relevance maxim)、和方式准则(Manner maxim)而造成的。事实上,许多修辞格,如反语、隐喻和夸张等,都是由于违反这些准则而产生的特殊交际效果。格氏认为,当说话人在字面上违反合作原则时,听话人会假设违反原则者在根本上仍在遵循原则行事;既然他是有意合作,他却违反准则并让听话者注意到他违反了准则,那么他一定是在传递一些符合原则的特殊信息,由此便产生了会话含义。同样的道理,蔑视词的意义规约,使用与语义韵相冲对的词组成异常搭配,也是对语义韵的利用,以期产生特殊的交际效果。典型搭配是为本族语者接受了的语言使用规约之一。人们在长期的语言使用实践中,都有意识地或无意识地遵循规约,选择典型搭配来实现意义。然而,当一个本族语者对规约明知故犯,有意地蔑视规约,使用语义韵相互冲突的词项来组成异常搭配时,他就破坏了搭配和谐。这种情况下,本族语者一般不会认为他的语言知识贫乏或语言能力差,以至不会使用正确的词语搭配,而会认为,这个人是知道正确的或典型搭配的使用惯例的,他之所以违反规约,一定是另有它图,旨在表达某种特殊意义,如正话反说,借事喻人等等。这样就创造了特殊的交际效果,产生了修辞学上所谓的反语、隐喻、夸张等修辞格。

下面,我们讨论几例本族语者利用语义韵创造反语修辞效果的情况。

洛悟(见 Baker & Francis 1993:157－174)专门研究了本族语者利用语义韵创造反语修辞效果的情况。他提到了米勒(Arther Miller)的名著 *Death of a Salesman* 中的一段对话：

Happy：(*trying to quiet Willy*) Hey, pop, come on now ...

Willy: (*continuing over Happy's line*) They laugh at me, heh? Go to Filene's, go to the hub, go to Slattery's, Boston. Call out the name Willy Loman and see what happens! Big shot!

照前面的分析,*happen* 一词具有强烈的消极语义韵,与其搭配的主语以及其他有关搭配词大都有消极涵义,指那些"不愉快的事"。在 *happen* 的语境内,人们会期待消极词语的出现。可是在这段对话里,紧随 *what happens* 之后的却是 *big shot*(大人物,大亨)这一词组。这就造成了语义韵冲突,破坏了语境中的搭配和谐。而 Willy 之所以这样做,正是要利用 *happen* 一词的消极语义韵,违反常规,创造一种反语和幽默的修辞效果。JDEST 语料库、LOB 语料库和 BROWN 语料库中,也不乏利用语义韵以期创造特殊交际效果的实例:

(1) *Cause the spectacular effects*

1.1.1 节中 *cause* 一词索引的第 27 行里有 *cause the more spectacular effects* 这一搭配。根据前面的讨论,在 V + N 类联接内,N 多为表示专业领域内人所不欲的"失败"、"事故"和"故障"等意义的词项,但这里却用了"壮观的效果"这个具有积极涵义的词组。它显然是一例违反常规的用法。但,该索引的扩展语境会显示出更多的信息:

... eradication of life within the river. One can therefore consider pollutants as being oxygen-depleting or directly toxic to the river fauna. Discharge of these types certainly cause the more spectacular effects on the river system. (JDEST Data)

很明显,这里的话题是河水污染。河里的生物可能会灭绝,污染物

可能会耗尽氧气，或者直接毒害河里的动物。向河内排放这类污染物当然会导致严重后果。然而，“河水受污染”、“动物被毒害”、“生命灭绝”这些地地道道的灾难性后果却被说成是“壮观景象”，其反语、讽刺和警示的特殊交际效果是不言而喻的！

(2) *Cause of conscience*

> He had a large sense of drama. He dramatized himself, and he dramatized the world of nature. Though others have been imprisoned for cause of conscience, he is remembered as the one who spent a night in jail for refusal to pay taxes to support ...(JDEST Data)

根据前面的分析，*cause* 一词在 N + PREP + N 类联接内也具有强烈的消极语义韵，介词 of 与 for 后面的名词搭配词多为 *death*、*trouble*、*anxiety* 等“消极词”。而这一段里却出现了 *cause of conscience* 这一说法，明显有违惯例。然而，分析整个语境，作者用此异常搭配的特殊用意便不难理解了。在这段文字里，作者调用了一系列词语手段对 *he* 进行讽刺、调侃和挖苦：“he”具有极强的戏剧意识，使自己的生活戏剧化，也使世界戏剧化，当别人由于良心的缘故而被监禁时，*he* 却因拒绝纳税而在大牢里关了一夜。在语境中，*imprisoned* 和 *jail* 等词的语义特点是和 *cause* 一词的消极语义韵一致的，但 *drama* 和 *dramatized*，*conscience* 等词却和语义韵冲突，造成搭配不和谐。当两类词共现于同一语境时，作者欲创造的“讽刺、调侃和挖苦”等修辞效果便跃然纸上了。

(3) *Utterly fearless*

BROWN 语料库中有这样一段话：

> The fact that she was missing assumed a very sinister as-

pect. Chinese Lily was quite capable of having an European girl kidnapped if she wanted to do so. She was utterly fearless, and had been too long associated with the underworld not to know exactly how to set about it. (Brown Data)

utterly 一词具有消极语义韵,在 ADV + ADJ 类联接中,它的形容词搭配词也多为消极词,*utterly fearless* 这一搭配也有违惯例。在上述语境中,*sinister*、*kidnapped*、*the underworld* 等词与 *utterly* 具有相似的语义特点,它们的共现可谓一种“和谐”共处。但 *capable of* 和 *fearless* 明显属于异类。两类词共现同一语境,必然造成语义韵冲突。但,这种冲突正是为了突出下列特殊交际效果: *she* 这个姑娘失踪了,这是个不祥之兆;因为一个叫 *Chinese Lily* 的人对绑架(kidnap)欧洲姑娘这类事很能干(capable of),此人与下流社会(the underworld)过从已久,完全是无所畏惧的(utterly fearless)! 这种冲突,这些异常搭配的使用,既反讽了 *Chinese Lily* 这个人缺乏良知,胆大妄为的丑恶本性,又强烈地突出了 *she* 已被绑架的可能性,表达了作者的判断和推测。

(4) *Incidence of skills*

incidence 具有明显的消极语义韵,在 N + PREP + N 类联接内,介词“of”后面的搭配词多为表示各个专业领域里“坏事情”的词。但 JDEST 语料库中却有“incidence of skills”这一搭配,其扩展语境如下:

The labor force is generally deficient in skills and, as in most Latin American countries, the incidence of skills is lowest in agriculture. In particular, skills are few among the subsistence farmers, who are virtually without education ... (JDEST Data)

在语境中，*deficient*、*subsistence farmers*、*without education* 这些词语可以说是和*incidence* 的消极语义韵一致的，而 *incidence of skills* 却似乎是个异常搭配。但，作者显然是为了突出这个国家"劳动力缺乏技能"这一事实，才如此进行词语组合的：先提供劳动力技能缺乏这一语篇信息，后又强调勉强维持生计的农民都没有受过教育，尤其缺乏技能，而在中间使用 *incidence of skills* 这一异常搭配，使其醒目和引人注意，从而突出农业人口技能缺乏这一严峻事实。严格说来，这一异常搭配的使用并非旨在反讽。但其修辞上的强调、突出作用却毋庸置疑。

总之，使用异常搭配，利用语义韵冲突创造特殊的语言交际效果是自然语言使用中常见的现象，无论在普通英语文本里，还是专业英语文本里，都时有发生。这是本族语者凭借丰富的语言、文化知识与语言使用经验，驾轻就熟地创造性使用语言的结果。从语义韵的角度，通过分析共现于同一语境中的有关词项的语义特点，无疑可以开辟一个新途径来理解英语的修辞手段与效果。然而，必须指出，语义韵的利用主要是本族语者的事。一般 EFL 学习者应首先学习和使用典型搭配，从而能恰当而地道地表达意义，而语义韵的利用则是更高一层的要求。

结论与启示

本章以 JDEST 语料库和其他有关语料库的证据为依据，对专业文本里的语义韵现象进行了探讨。这只是一个小规模的尝试性研究。尽管如此，研究还是清楚地表明，科技和学术研究文章中也存在着语义韵现象。研究还表明，本族语者在一定的语言使用环境下，为实现某种特殊交际效果，会违反规约，使用异常搭配，利用语义韵冲突来达到交际目的。

语义韵研究对语言学与应用语言学研究有着重要的启示与意

义。它再次证明，词语的使用与语境是不可分离的。在语义韵中，典型搭配词营造起了一种特有的语义氛围。这种语义氛围影响整个语境；它弥漫到整个跨距，渗入整个类联接，甚至超越句界。位于语境内的词项都在不同程度上受到特有语义氛围的传染。它说明，词项的真正意义就在语境之中，存在于与之结伴的别的词项之中；研究词义必须从研究语境、研究典型的词语搭配开始。那种认为词项具有单独的自成一体的意义的观点是不符合语言使用实际的。语义韵的发现还再一次证明了语言交际的"意义驱动"本质：意义制约着词语与句法结构的选择，一切语言形式的选择都要受制于意义实现。它向传统语言描述体系中将句法置于首要地位、词语置于次要地位的做法提出了挑战，并给人们以重要启示：词项并不仅仅是实现一定句法结构的填充材料。它有自己的行为模式，在实现交际意义中起着极为重要的作用。词语行为研究应当成为语言研究的出发点并居于中心地位。

语义韵可能与语言社团的社会文化因素以及心理因素等有着密切关系。从某种意义上说，它是因语言而异的。在英语里，*cause* 一词具有强烈的消极语义韵，英语本族语者常用 *the cause of failure the cause of trouble* 等说法。与此相反，汉语里却常用"成功的原因"、"经济飞速发展的原因"、"取得进步的原因"。如果将汉语的表达方法直译为 * *the cause of success*，* *the cause of economic development*，* *the cause of progress*，则完全违反了 *cause* 一词的典型搭配行为，破坏其一般的消极语义韵。同样，如果将 *happen* 一词作为 *occur* 和 *take place* 的同义词教给 EFL 学生，也是不确切的。因此，语义韵的研究会对 EFL 的词汇教学起着十分有价值的指导与启示作用。随着研究的进一步深入，EFL 教学实践者将获得这方面大量有价值的信息，帮助学生学习到地道、自然的英语。

语义韵是一种独特的词语搭配现象。它不同于传统词汇学所描述的褒义词和贬义词的概念。在词汇学的描述体系中，一个褒义词本身具有明显的褒扬的或积极的意义，如 *individualism*（自由主义）；一个贬义词也明显地有其消极意义，如 *villain*（恶棍）。但它

们的搭配行为却不见得都是一些"积极词"或"消极词"。然而，就语义韵这种现象而言，很难说节点词本身有多少明显的积极意义或消极意义，比如 *cause*，*incidence* 等词，而是它们的搭配词常具有一定的积极或消极语义特点。语义韵的形成可能与词项意义的褒扬与贬抑变化有关，但这仍需进一步研究。

语义韵的研究才刚刚起步，目前所发现的语义韵现象可能只是真实语言使用中的冰山一角。无论对语言学，还是对应用语言学的理论与实践，语义韵研究都具有广阔前景和重要的意义。但是，研究必须是数据驱动的。惟有依据语料库的强大证据，通过对数据进行统计、分析与概括，才能科学地描述词项可能有的语义韵。当然，本族语者对词项的语义韵会有某种意识或直觉。但总体而言，任何人的直觉对语言研究都是极其有限和靠不住的。再之，语义韵研究与描述应采用概率的方法(Halliday 1991)，因为词语行为不是绝对的和死板的，除了语言使用的创造性因素外，还有个人用语特点和专业用语特点等因素。所以，概率的方法是一种较为科学的方法。



第九章 千年之际展望语料库语言学*

1 引语

本文对目前语料库建立中的实践与重点事宜以及语料库利用技术进行评述。评述主要依据欧洲最近几年的经验——在欧洲，自然语言处理的传统已经开始与语料库语言学融合。它表明，语料库语言学已经取得了很大成就，但在很多业已确立的实践中，甚至在其理论基础方面，仍缺乏长期的稳定方针。欧洲委员会成立了一个机构——语言工程标准专家咨询组(简称 EAGLES)，来研讨此类问题。该机构有个部门专门负责语料库研究。本文就是根据广泛而详尽的语言工程标准专家组的报告写成的。这些报告对语料库语言学工作者来说特别及时和有益。专家组成员试图建立一些方针来协调方方面面的活动，结果是确认了：

- 大量有益的实践活动
- 许许多多的共同目的
- 最近达成的关于许多问题的共识。这些问题曾在过去引起争论。尽管有强烈的融合迹象，但情况也很明白：在语料库研究的许多领域，情势仍很不稳定，难以制定和实行明确与严谨的标准。其原因是双重的：
- 语料库的概念仍在发展。语料库的基本涵盖范围正在扩大。语料库可能使用的结构正趋于复杂。在这种情况下，进行硬性规定是不适宜的。

---

* 本文为辛克莱(J. Sinclair)教授所撰写，经作者授权同意收入本书作为第九章。

- 大多数学者用于解释和描述语料的理论框架已经过时，且决不可能有效地适应此项工作。

因此，几年后的情景很可能会大不相同。语料库将由自动化程序建立；直接对语言输入进行操作的完全自动化软件将使用户能够获得语料。语料库在急速地变大，需要开发可靠的自动化程序。无论其目的如何，大规模语料库的优点已显而易见。值得庆幸的是，急速增多的大量可获得的电子出版资料已能满足这种需求，用于处理资料的硬件与软件也日益增多。

2 分类

语料库的分类是从一些主要的一般定义开始的。分类提出了重要的问题，随着语料库信息的使用，需要加以解决。

或许，最重要的一点是人们认识到了，目前所进行的建设一般语料库资源的活动正在为行将到来的"监控语料库"(monitor corpora)提供建设基础。到目前为止，语料库大体上由单一时间段里选取的文本集构成；时间并不是变量。现在，文本已可以大批获得，定期取样并建立时间变量已成为可能。因此，需要新的软件工具和操作办法来管理新的语料库。而且，现在开始谋划并非为时过早。

为语料库选材而进行的文本分类已意味深长地摆脱了早期暂定的标准，力图使其尽可能客观。已将外部标准(来自使用文本的社会)与内部标准(依文本的语言特点而定的语言学标准)分开；前者成为分类的基础。内部标准——诸如题材与文体，目前较难加以确定。

在早期的语料库分类活动中，人们曾经对文本题材这个含糊不明的概念给予了注意，以便平衡地选取题材和平衡地覆盖词汇。然而，这种做法是相当不科学的。菲利普(Philip)等人的研究表明，题材的内在表现是复杂的。随着语料库日益多语种和多文化，凭直

觉划分范畴受文化制约这一点便显而易见了；一人之癖好，可为他人之职业，也可为另一人禁为之事。

未来几年的目标是取得更大程度上的系统完善，因为我们要使语料库更加可靠。为了指导语料库成分的选择，我们有必要了解某些语言变体是如何取得“标准”地位的，各类语言在社会中的“渗透”已被确定为优先研究对象。除此之外，研究文本的内部结构以发现有实用价值的标准，也是极为重要的事。

3 标记与注释

最近几年人们对格式标准化做了大量工作；格式制约着外来材料，包括版面布局、解释或文件信息（篇首信息）等插入数字字母字符串的方式。这方面取得了很多一致性意见，我们应当继续制定出指导方针，以妥善放置这种复杂的外来材料。

重要的一点是我们现在做出的建议必须被视为是临时性的，因为文本本身的性质在发生变化。SGML（标准普通标注语言）和TEL（文本编码方法）是早期为防止书面文本的格式与版面在变为电子形式时丢失而设计的办法。其基本策略具有双重内容：在连续文本中附加一些编码；在篇首处加一段关于文本的信息。到目前为止，这两种活动大体靠人工完成。只是到最近，市场上才出现一些电子辅助手段。与此同时，越来越多的文本由电子程序写成，方法日益混乱纷繁。

有几点现已清楚。第一，从现在起，所有的格式注释都将由自动化手段进行。在这种情况下，我们目前的许多问题及其探询都将不复存在。如克里尔（J.Clear[2]）诙谐地指出的那样，目前情势下的危险是：标记部分极易“溜入”解释部分。

第二，未来的注释部分将与文本本身隔开。情况似乎是，来自各方面的压力会要求建立一种类似“最小文本”的国际标准，用作由

社会与法律确认的对文本管理的基本水准。因此,所要求的应是"分离"而不是"是否可以分离"。在这种背景下,语言工程标准专家组关于篇首信息(现在通常置于文本之前)应当置于单独文件中的建议是应当受到欢迎的。(请进一步参阅下文的"口语语料库"一节中有关"合法性"的段落。)

第三,计算机日益增大的能量和人类对语言的进一步理解将使用户进行文本预处理和文本描述时采用的两种明显不同的方法趋于一致。这些方法可称为"标识方法"(annotation approach)与"实时方法"(real time approach)。如果用标识方法,文本提供者试图去猜测用户可能需要的内容,因而花费大量资源对文本添加种种注释。由于经过各种各样的方法"校正"过,文本一般都很整齐,而且其长度可能增加几倍。用"实时方法"则对文本不加改动或增添,但要开发软件来实时提供由标识方法所提供的那类信息。

过去几年间,这些都似乎受到过激烈反对,尤其是当"注释方法"还不可避免地牵涉到人工介入。现在可以看出,两种方法仅在低层面策略上有所不同,也就是说,是根据要求实时分析还是预先分析。显而易见,这是个个人使用的问题,而不是其职责为提供一般资源的研究人员的关切所在。

可以看出,我们的工作涉及三个变量:

- 增加的内容是由人工来做还是靠自动程序。
- "未标注文本"[3] 是否保存;非文本信息是作为注释保存在连续文本中,还是保存在另外的相关数据流中。
- 解释性工作是预先做好还是应需而做("实时")。

然而,这三种对比并非独立的变量,因为几种变量结合一体的情况会出现。比如,"人工+未标注文本+预先"。过去十年间的常规选择是"人工+注释文本+预先"。而且,由于人工注释仅适用于某一特定文本,也就没有必要将其与文本分离。由于一些庞大的注释项目已逾数百万字,注释文本之臃肿成了严重问题:降低和限制了使用,几乎只有最刻意执著者才会去使用。然而,资助者不再愿为这种活动付钱。再之,可供使用的语料库的规模增大得如此之

快,不久,人工注释的可能性将不复存在。

我认为,未来的十年中,"自动、未标注文本"的选择将处于主导地位。道路已经畅通,人们可以使用成套软件,以实时方式,迅速又高效地做大部分工作。对于那些可预见的并经常需要某些注释的项目,在整个语料库中使用工具或许是值得的;可将注释结果储存在数据流中,需要时则与连续文本合并一体。分析采用自动手段时,"预先"与"实时"间就不存在原则上的差异——差异仅在于策略。对提供一般资源而言,由于要求规模与灵活性,"自动+未标注文本+实时"之模式很可能会普遍盛行。

4 工具

从上述论点可以看出,重点将由注释转向工具。回过头来看,注释可被视为一个时期内出现的一个过渡阶段。那时,我们的语言学技能与计算设施都还不能足以恰当地从事工作。这种不足导致了几个中间阶段的产生,充当生文本与分析间的缓冲区。使人为之茫然的各种各样的工作平台、预处理器等被广泛使用。还有可能建立一种辅助系统来评估工具的标准。为减少这种毫无必要的复杂行为,所有的工具可参照上述建议的"最小文本"标准相互联系起来。一种工具既可用来确保接受最小文本的输入,又可用来确保接受另一具体工具的输出。工具的输出当然要按照目前的惯例加以规定。

必须再次说明,这种立场决不是要反对为某个特定的使用目的而准备标识文本。这纯粹是使用项目管理上的策略问题。注释文本从来就没有一般资源理想,因此已没有必要再使用它们。在未标注文本模式下标识时,恢复最小文本并将其与自动工具分别保存将是微不足道的简单过程。

语料库与相关的电子资料已经有了信息交换所。例如,美国的语言数据联合会和欧洲语言资源协会(ELRA)。这些组织的目的

就是从出品人那里收集材料，使其能为用户所用。它们对资料到达时所处的状态影响极小。但是，如果它们能一直确保最小文本版本并建立其目录，就将对语料库的发展与其结构之清晰作出显著贡献。

该办法旨在协调各种实践活动，仅给开发者略微增加了负担，不会阻碍进步。

5　结构分析

这里可以最为清晰地看出，我们对于和语料库建设相关的自然语言处理的观念正处于转折点。关于目前工作现状的报告显示了语料分析方面所取得的巨大成就与成熟程度。对几种语言进行词类赋码（称为词法句法分析）已相当准确，至少，有一种称为制约语法（constraint grammar）的句法分析方法已达到几乎自动化的程度，并相当准确。其他各类分析——语义、语用和语篇分析仍处于初期和不确定的发展阶段。

现在，人们认识到了转用自动分析方法的必要性。或许在下一步的句法研究与语义分析等工作中，重点将会是尽快地设计一些完全自动化的工具。（我们同样希望，某些急迫的使用项目将不得不做出临时和非正式的分析，但这些分析最好不要与提供一般资源相混淆。）

尽管结构分析的方法多种多样，各自明显不同，但重要的是，这些方法大都来自一个源头。从实质上说，它们都产生于或者说紧密联系于在语料库问世之前久已存在了的“公认的语法”（consensus grammar）。这种语法的分类范畴及其重要做法几乎没有受到过严肃认真的挑战。务必注意的是，计算机并没有被用来对这些语法范畴进行系统的检查，而是视其为当然正确，将其强加给数据；当其明显地不适合时，才作一些最小的调整。

然而，语料库是允许对各种各样的假设进行验证的，无论它们是正统学说还是异端学说。学者们在运用"公认的语法"研究文本语料时遇到的困难表明，"公认的语法"那些假说可能不是最好的，至少不是最合适的。的确，关于人类语言分析的许多受到严密保护的假想是可疑的，很可能要从根本上修正。

因此，试图将目前的"发展现状"惯例作为未来的标准是不恰当的。语料库进行的结构分析，其目标极不明确，也没有进行过明晰的表述；它们被假定为是对语料库问世前的分析目标的映射。下一步的最佳战略是对目标进行探讨，进行重新解释，据我们所知——目前关于结构的假想都不可信。

6 基于语料库与语料库驱动

托格尼尼—布纳里(E. Tognini-Bonelli[4])区分了运用语料进行语言学研究时两种截然不同的方法，并澄清了目前的工作现状。基于语料库(corpus-based)的研究就是使用语料来充实分析，改进分类范畴，填补类别成员，消除歧义和描述上的问题，并且提供大量的真实例证，坚实有力地论述自己的观点。语言学家驾驭研究，证据被查询，描述得到改善。没有迹象表明，由于查询了语料，可能要进行任何反思；因为语料库仅仅被视为一大批熟悉的信息。描述分类不可能改变，理论本身戒备森严，反对任何可能来自语料的反拨；因为理论已事先宣布，用法的描述不能增添意义。

进行语料库驱动(corpus-driven)式的研究，研究者对理论的能力及其对所有将被语料库研究证明为重要的结构模式的描述解释远没有那么大的把握；这些结构在许多情况下都传载意义。如果采取这样的态度，即，既然人类意识的正常水平仍不能捕捉许多的语言组织机制，(如果不是大多数机制的话，)我们就有可能发现很多由大体上基于直觉的理论所绝对难以预见的东西。要期待发现

新的分类范畴，丢掉老的区别；要准备以尽可能开放的胸襟去观察语料——当然，不是完全开放，因为甚至观察也有待于基于某种理论的解释，无论该理论是如何的模糊不明；但是期待的情况是，现行理论已侵入的地方是最难以新的方式描述的。

当研究导致的新理论和描述从老理论的废墟上诞生时，可能被发现与先前者相似。没有充足的理由认为，大多数语言学家在大部分时间里的研究都是错的。例如，主张必须根据词汇在语料里的相似行为模式重新思考词类划分时，语料库驱动的语言学家并不否认名词和动词的"存在"。极有可能，对许多语言来说，此类主要的分类范畴将支撑未来的描述。但是，不同于今天，它们将以科学严谨的方式加以界定，每一个词类的成员都是可列举的，词类间的重叠将精确地加以区分。

7 多语语料库

近几年来，翻译语料库，或者叫"平行"语料库，十分流行，尽管人们担忧文本译为目的语后可能被曲解。平行语料库的流行是由于设计自动对列软件的工作似乎取得了初步成功。只要译文保持着易于计算机辨认的边界，比如段落和句子界限，几个对列软件就可以产生较好的结果，虽然常常需要人工做一部分预处理。在现实的使用项目中，比如最近报道的 TELRI 项目，由于大多数译文与原文相比不很整齐，现行软件的用途便极为有限。

尽管如此，使用外部与内部标准相结合的词组水平上的自动对列(程序)的前景仍相当乐观。如果情况如此，我们可以预见，平行语料库将被用作未来词典编撰的资料源。作为译员知识的贮存器，它们将形成一种重要的资源。况且，已有大量合适的语料库材料在全世界的多语社区出版。情况至此，就像一般语料库那样，它们有可能变得太大，难以进行人工处理。规模对于捕捉译文的细微差别

有巨大价值,但需要开发完全自动化的软件。

然而必须记住的是,这样的语料库构不成一般意义上的语言资源,因为它们并不是有关语言的代表性样品。不可避免,它们只出现在人类行为的某些领域,主要是较为正式的政治、技术、商务和法律领域,也涉及一些新闻;因此,包括非正式的、即兴的和口语材料的普通语言的双语语料库是不可能建立的,如果没有专门用于该语料库的大量翻译努力。此外,人们普遍认为,翻译语言的特点和翻译有关,但不反映语言的特点。所以,由翻译文本组成的语料库不是整体语言的良好范例。

那末,我们可以预见,重点将转移到由欧盟 PAROLE 项目首创的"比较语料库"。在"比较语料库"里,每一种语言的文本通过比较有关社区的语言行为来选取。现已清楚,语料库分类方面的某些普通标准是由具体的文化限定的。那么,上述所建议的外部标准研究看来对多语和单语工作都同样需要。

8 口语

口语语料库与书面语语料库应当一起发展,这一点得到普遍同意。由于许多国家学术研究的传统所致,语音研究已变得有点专门化,因而需要自己的材料及材料获取方法。那么,口语语料库可能是两种极为不同的事物。为语音或"言语"研究而建立的录音数据库,或者用于一般语言学研究的译成文字的口语数据库。当然,有许多中间的可能性,因为言语数据可以用种种方式译写。而且,普通语料库中的口语材料惯例并不像书面语材料那样刻板。

这两种传统现在并肩工作;每个传统的发展都会产生共享语料的机会。在欧共体内部,已决定将书面语语料库与口语语料库资源的开发分开,将口语资料的职责交给了言语科学家。该决定的作出很可能是基于经济原因,从专业角度看是有问题的;但存在的危

险是,在大型参考语料库中,口语没有给予应有的优先考虑。这种做法与语料库用户所持的观点相冲突,他们强烈地认为语料库中的口语成分特别重要。

协调口语与书面语材料收集的关键是正字法;无论语言学分析什么,正规的拼写与发音应当保持。如通常情况那样,如果要提供进一步的精密注释,它们应当一直被妥善地置于相对于连续文本的合适位置,而不是将其搞得模糊不清。

语言的书面形式的一般正字法有其特性,却往往被受过高级训练的专门人员忽视;这就是易读性。人们可以毫不费力地读懂。每个在语言文本资源领域工作的人员都应当忠实地保持这个特性,无论改变正字记录是多么便利于计算目的。

人们还指出,秘书训练可使很多人能将口语文本转写为一般的文字,这使得建立口语语料库的高额成本控制在一定范围之内。任何干扰惯例的作法,甚至表面上微不足道的增添,如记录停顿,都证明会大大减慢转写速度。这样,就要求打字员成为某种形式的语音工作人员,而且发现,这样的决定使得打字员工作与常规的秘书抄写有着重要不同。

数字化录音和声波分析自动化工具的出现使我们愈加可能发现言语与文字间的进一步的相关性。由于其实用性,这个及时到来的研究方向应受到鼓励。对标点符号与韵律间的相互预见性进行的研究尤为重要并大有希望。这方面的成功会对言语识别作出实质性的贡献。

9　词典

在未来几年中,极有可能在词典领域里进行彻底检查。用于设计今天机器词典的语言理论十分适合于捕捉科技术语的明显特征,因为科技术语的意义截然分明,而且相对稳定。许多支持这类词典

的论点都论及技术文献的信息检索——所基于的假设是,交际活动中的信息内容是稳定的。

但是,这些词典却难以处理普通词汇的使用模式。当用于处理语言的一般词汇时,词典的构建成本就昂贵且操作性能极差,因为这种模式不适用于数据——这是基于语料库的语言学家难以敏锐察觉语料特点的一个有力例证。

未来十年中,建立新的词典模式极为必要。这些新模式要用来处理搭配、语法共选、语义和谐与态度意义(attitudinal meaning)等方面的信息,——这些都是语料库研究要首先研究的几类结构模式。这些内容与基于属一种分类和抽象的语义特征的词典几乎没有多大关系。但是,它们产生于对用法的研究,处理的是意义,因而与语言工程的运用明显相关。

由于对语料库问世前模式(pre-corpus model)的词典进行了大量投资,我们可以预见该领域会出现反对改变的强大力量,尤其是所要求的调整是相当根本性的。这些改变包括下列内容:

- 新词典的特征将是词汇语法式的(lexico-grammatical),不首先对句法和词汇进行区分,而是将句法说明视为词汇描述中的较为抽象的表述。
- 语料库研究表明,词或词组的意义对选择与其共处同一环境的其他词与词组有着深刻的影响;这种相互依赖的关系从未被正式表述过,但它的重要性十分明显,因而必须纳入词典设计中。
- 将词作为意义单位是低效率的做法。目前,大多数词书,如词典,都是围绕词来安排意义;在绝对不可避免的情况下,才引用词组作为独立的意义表达单位。新的词典很可能基本上以词群为单位,伴之以大量的内部结构变化。

简言之,与目前的模式不同,组合关系轴上进行的连续选择的全部融合力要纳入词典。目前的模式承认诸如配价这样的组合选择的重要性,但将其置于完全封闭的语法部分里。

10　总述

本文对发展趋势进行的总结使我们对即将到来的变化作了几点预测。这些变化共有的特征将使其难以被语言文本分析研究者们接受。原因是:新的发展所依赖的基本语言模式要比老模式简单。最大的阻力就在于此。要使虔诚敬业的学者们相信,一个学科长期发展起来的复杂而又深奥的机制是不必要的,这实在太难;而且,人们会担心自己的声誉与未来工作,因为由此而显露端倪的是如下信息:

- 语料库中的文本最好不加标识
- 工具应当对生文本进行操作或者操作别的工具的输出
- 在语料准备过程中或工具使用中不应当再有人工干预
- 目前的词典模式将被限制在术语领域内使用

由于语料库本身的性质,所有这些变化虽然可被理解为不可避免,但却形成了可怕的理念问题。

语料库语言学之所以出现这样的情势,是因为它还没有系统地阐述过自己的理论和描述方法。被主流语言学冷落了三十年后,语料库开始被计算语言学家视为有用的资源,因为这些计算语言学家开发了复杂的描述程序但却发现难以用之处理普通文本。他们的理论与描述方法是不加批判地引进的,后来花了巨大努力利用从语料库搜集来的信息去改变程序的功能。这就是基于语料库的语言学(corpus-based linguistics)的主要类别之一。

现在,情况已趋于明白,问题不仅仅是调整现有的方法,而是进行根本上的变革。支撑这些模式的原有理论立场是厌恶数据的。它认为,以用法为据进行语言描述没有任何益处。它主张,语法体系的建立要由内省检验,而不是靠外部证据。在理论的基本设计中,实际中使用的言语和文字尤其不被注意。那么,出现这些情况

就不足为奇了——人们在描述上碰到的问题，常常似乎难以逾越；而很少有人接受可能的简单解决办法。

一些不赞成多数人信奉的心灵主义思想的学者采取了中间道路，试图复制所谓"公认的"语法。这类描述很适合于能流利使用相关语言的人们，但却难以转用于对语料库进行自动描述。其实际问题与心灵主义者碰到的问题没有很大不同——从本质上说，这些描述有两个方面与事实不符：首先，它们不够精确；其次，它们忽视如果将其包括进去就会极大地提高描述精确性的至关重要的语言范型。这是另一个主要类别的基于语料库的语言学。

重建理念的问题是严肃的。为未来建设语言资源的学者们承担着严峻的职责。千年之交即将来临，进行变革和再生长的时代即将来临。巨大的优势是问题已被明确表述。主张做好变革的准备并留有充足余地的人们认为，借助计算机来理解语言的潜力是十分巨大的，不应当被语料库问世前的方法惯例阻碍，即使实际上发生的永久性变化可能不大（我们真的不知道！）。

（卫乃兴译）

注　释

1. 辛克莱(J.Sinclair)：本世纪杰出的语言学家，伯明翰学派的代表人物，长期从事语料库语言学研究，是COBUILD项目的领导者和组织者，COBUILD系列著作的主编。本文是辛克莱教授应杨惠中教授的邀请而撰写的一篇评述性文章。文章介绍语料库语言学发展的总体情况，总结其一般的原则与方针，并预测下个世纪里语料库语言学发展的趋势。本文由卫乃兴翻译为中文。
2. 克里尔(J.Clear)：伯明翰大学语料库语言学家，重要论文有"Trawling the language: 'monitor corpora'", "From Firth principles: computational tools for the study of collocation", "The state of the art in International Corpus Linguistics","Grammar and nonsense: or syntax and word senses", "Technical implications of multilingual corpus lexicography"等。

3. 未标注文本(clean text)，指该文本未加任何标注、赋码等，保持其原始状态。
4. 托格尼尼—布纳里(Elena Tognini-Bonelli)：伯明翰大学语料库语言学家，主编有 *Terms in Texts*，*Text and Technology: Studies in Honour of John Sinclair* 等语料库语言学方面的系列著作。



附：第九章原文

Corpus Linguistics at the Millennium

© John Sinclair 1997

This is a review of current practices and priorities in language corpus provision, and the techniques for exploiting the corpora. It is based on experience over the last few years, largely in Europe, as the traditions of Natural Language Processing and Corpus Linguistics have begun to merge. It shows that a lot has been achieved, but that there is no long- term security in much of the established practice, and even in its theoretical foundations.

A body set up by the European Commission to consider this topic among others (EAGLES — Expert Advisory Group for Language Engineering Standards) had a section on corpora, and this paper is written in the light of the extensive EAGLES reports. (ref) The reports are particularly timely and helpful to those working in the field. The efforts of the group members to suggest guidelines for harmonising their activities have led to the identification of:

~ a great amount of good practice;
~ a considerable amount of common purpose;
~ a de facto consensus, recently arrived at, concerning many issues that have provoked argument in the past.

Despite this strong indication of convergence, it is also clear that in most areas of corpus work the situation is too unstable for explicit and rigorous standards to be formulated and applied. The reasons are twofold:

~ the notion of a corpus is still developing; the basic dimensions are enlarging, and the structural possibilities are becoming more sophisticated. In such circumstances it would be inappropriate to apply inflexible rules.
~ the theoretical frameworks with which most scholars interpret and describe corpus evidence are out-of-date, and indeed were never likely to be well-oriented to the job.

It is therefore likely that the picture will be radically different in a few years time. Corpora will be built up by automatic procedures, and the evidence will be made available to users by fully automatic software operating directly on the language input. Corpora are rapidly getting so large that the reliability and replicability of automatic procedures is now required, and the virtues of size, almost regardless of the purpose, are becoming obvious. This need is thankfully matched in most cases by rapid increases in the amount of data available in electronic form and the growing provision of software and hardware that can handle it.

TYPOLOGY

Work on the classification of corpora has begun with some broad general definitions and raised issues that are important and need to be worked out as information about corpora becomes available.

Perhaps the most significant point is the recognition that the present activity in the provision of generic corpus resources is being done to provide a basis on which can be built the "monitor corpora" that will shortly follow. Corpora until now have largely consisted of a collection of texts taken from a single time-band; time has not been a variable. Now text is available in sufficient quantities to make it possible to take samples at regular intervals, and build up a time dimension. New software tools and operational practices will be needed to handle the new corpora, and it is not too early to start planning them.

The classification of texts for inclusion in corpora has moved significantly from early criteria, which were ad hoc, to attempts to be as objective as possible. External criteria (drawn from the society that uses the texts) are separated from internal criteria (drawn from the linguistic preferences in the language of the texts), with the former forming the foundation of the classification. Internal criteria — such areas as topic and style — are more difficult to identify at present.

In the early days of corpus classification, some attention was paid to a very vague notion of what a text was about, so that, hopefully, a balanced selection of topics would lead to the emergence of a balanced vocabulary. This was, however, quite unscientific, and the work of researchers such as Phillips (ref) showed that the internal representation of topic is complex. As corpus work becomes more multilingual and multicultural, it is soon apparent that the intuitive categories are culture-bound; what is one man's hobby is another's career, and another's crime.

Greater sophistication is aimed at in the coming years, as we seek to make corpora ever more reliable. In order to guide the selection of corpus components, we need to know more about the way certain varieties of a language achieve "model" status; research into the "penetration" of various types of language in society is identified as a priority. Also major research into the internal patterns of texts to find useful criteria is also most important.

MARK-UP & ANNOTATION

There has been a great deal of work in recent years towards standardising the formats by which the stream of alphanumeric characters is interrupted by extraneous material, whether about layout, or interpretation, or documentary information (called "header"). A large body of agreement exists, and we can move on to frame guidelines for deploying this sophisticat-

ed but intrusive material.

The central point is that our recommendations of today must be seen as highly provisional, because the nature of text itself is changing. SGML (the Standard General Mark-up Language) and TEI (the Text Encoding Initiative) arose from early attempts to keep aspects of the format and layout of a printed text from being lost as it was put into electronic form. The basic tactical decisions were twofold: to include additional codes in the running text, and to add a section at the beginning which gave information about the text.

Both of these operations have largely been done by hand up to date, with some electronic aids coming on the market fairly recently. Meanwhile, more and more text is composed by electronic process, in a growing chaos of verbose conventions.

Several points are now clear; first that any and all format annotations will be automatic from now on, and in that case many of our current problems and queries will just not arise as a result. The danger in the present situation, as Clear points out wittily (ref), is that mark-up slides too easily into interpretation.

Secondly, the annotations of the future will be kept separate from the text itself, and it seems likely that from various quarters pressure will arise for an international standard of something like "minimal text", used as a baseline for the legal and social identification and handling of texts. Hence separation, and not separability, will be called for. The recommen-

dation of EAGLES is that "header" information — the section that is nowadays usually placed at the front of the text — should instead be kept in a separate file. This suggestion is to be welcomed in this context. (See further reference to "legibility" in "Spoken Corpora" below.)

Thirdly, the increasing power of computers and human understanding of language will bring together two apparently contrasting approaches to the presentation and preprocessing of text for the user. These can be termed the "annotation" approach and the "real time" approach. In the annotation approach, the text provider spends a lot of resource adding all kinds of annotation to the text, trying to guess what the user is likely to need. The text is typically very tidy, having been "corrected" in various ways, and it may be several times its original length. In the "real time" approach, the text is not changed or added to, but software is developed that will provide virtually instant information of the kind carried by annotations, to be applied where needed.

Over the last few years these have seemed to be radically opposed, especially when the "annotation" approach was also associated with manual intervention. Now it can be seen that they only differ in low-level strategic terms, i.e. whether the analysis is actually performed on command or prepared in advance. Clearly this is a matter for individual applications, which are not the concern of those charged with the provision of generic resources.

We can thus see that there are three variables to work with:

~ whether the added work is manual or automatic
~ whether the "plain text" is preserved; whether non-text information is kept within the running text as annotation, or in a separate, linked data-stream
~ whether the interpretative work is done in advance or as the occasion arises ("real time")

These three contrasts are not independent variables, however, since several of the combinations are not likely to arise; "manual, plain text, in advance", for example. The regular choice of the last decade has been "manual, annotated, in advance", and since a manual annotation applies only to a particular text, there is little motivation to keep it separate from the text. With some big annotation projects running into many millions of words, the bulkiness of the annotated text is now a serious problem, slowing down and limiting access to all but the most determined; but sponsors are no longer as willing to pay for this activity, and the size of available corpora is rising so fast that soon the possibility of manual annotations will disappear.

With the "automatic, plain text" choice, which I believe will dominate the work of the next decade, the way is clear for the introduction of software suites which can be developed to be fast enough to do most of the jobs efficiently in real time. For applications that can predict very frequent need for certain annotations, it might be worthwhile to run the tool over the whole corpus, and store the results in a data stream that can be merged with the running text where necessary. There is no difference in principle between "in advance" and "real time"

when the analysis is automatic — only a difference in strategy. For generic resource provision, with its requirements of size and flexibility, the "automatic, plain text, real time" model is the most likely to prevail.

TOOLS

From the above argument, it can be seen that there will be a major shift in emphasis from annotation to tools. Annotation will be seen in retrospect as a transitional stage that arose at a time when neither our linguistic skills nor our computational facilities were adequate to do the jobs properly.

This inadequacy has led to a number of halfway stages being established, to act as buffers between the "raw" text and the analysis. A bewildering range of platforms, preprocessors, etc. are widely used, and secondary systems of evaluating the protocols for tools are in danger of being set up. To reduce this needless complexity, all tools could be related to each other with reference to the "minimal text" standard proposed above. A tool would either guarantee to accept the minimal text as input or the output of another (specified) tool. The output of tools would of course be specified according to current conventions.

It must be repeated that this position is in no way opposed to the preparation of annotated text for a particular application, which is purely a tactical matter for the applications project

management. Annotated texts were never ideal as generic resources, and it is no longer necessary to have them. Where texts have been annotated under a plain text model it will be a trivial procedure to restore the minimal text and present it and the (automatic) tool as separate items.

There are now clearing-houses for corpora and associated electronic material; for example the Linguistic Data Consortium in USA and ELRA(European Language Resources Association). Their objectives are to gather suitable material from producers and make it available to users; they can have little effect on the state the material when it arrives, but they will make a notable contribution to clarity and progress if they always ensure that there is a minimal text version, and build their catalogues around it.

This approach to the harmonisation of practices puts only minor burdens on developers, and does not impede progress.

STRUCTURAL ANALYSIS

Here is seen most clearly the turning point of our conceptualisation of natural language processing in relation to corpora. Reports on the present state of work suggest considerable achievement and maturity of analysis. Word-class tagging (called morphosyntactic) for several languages is now quite accurate, and at least one syntactic parsing strategy (constraint grammar) is almost automatic and reasonably accurate (ref). Other kinds of analysis — seman-

tic, pragmatic, discourse — are in very early and uncertain stages of development.

And the need to switch to automatic methods is now recognised, so presumably in the case of further syntactic work, and semantic, etc. analysis the emphasis will be on devising fully automatic tools as soon as possible. (Again it is to expected that some urgent applications will be forced to produce provisional and informal analyses, but these will not be confused with the supply of generic resources.)

Despite the apparent heterogeneity of approaches to structural analysis, it is important to remember that most of them come from just one source. Essentially they all are derived from or are closely related to "consensus grammar", dating from long before corpora became available. The categories and priorities of this grammar are hardly ever seriously challenged, and it is important to note that the computer is not being used systematically to check them, but takes them for granted and tries to impose them on the data, making minimal adjustments when they clearly do not fit.

A corpus, however, allows all kinds of hypotheses to be tried out, whether orthodox or heretical, and the difficulty that scholars have had in applying consensus grammar to text (documented in the extensive literature on this subject) evidence suggests that those of consensus grammar may not be among the best, or at least the best fitting. Indeed, many of the well protected assumptions of human language analysis are suspect, and probably due for radical revision.

It is therefore inappropriate to try to turn current "state-of-the-art" conventions into required standards for the future. The goals of the structural analysis of corpora are not at all clear nor explicitly set out; they are assumed to be a projection of the goals of pre-corpus analysis. The best next step will be to address the question of goals, and to reinterpret the goals in the knowledge that none of the existing assumptions about structure can be trusted.

CORPUS BASED AND CORPUS DRIVEN

The identification of two fairly distinct approaches to using corpus evidence in linguistics is made by Tognini-Bonelli (ref), and clarifies the present state of work. To be "corpus-based" is to use the evidence of corpora to inform your analysis, refine your categories, fill out class membership, resolve ambiguities and descriptive problems, and provide a fund of authentic examples from which you can choose those that make your points with greatest force. The linguist is in charge, the evidence is consulted, and description is improved. There is no suggestion that any rethinking might be required as a result of examining corpus evidence, because it is regarded as just a large amount of familiar information. Descriptive categories are not likely to be changed, and the theories themselves guard against any possible "backwash" effect from evidence, since the theories have pronounced in advance that the description of usage cannot add meaning.

To be "corpus-driven", you are much less sure of the ability of the theories and their descriptions to account for all the structural patterning that will be shown by corpus research to be important, and in many cases meaning-bearing. You adopt the attitude that since much, if not most, of the organising mechanism of language is below the normal level of human consciousness, the likelihood is that we are going to discover a lot of things that are in no way predicted by theories based largely on intuition. You expect to find new categories, to forget old distinctions; and you are prepared to look at the corpus data with as open a mind as possible — not, of course, totally open, because even observation is subject to interpretation based on some theory, however vague; but expecting that those places where existing theories are allowed to intrude will be the more difficult to describe in a new way.

The resultant theories and descriptions, when they arise from the ashes of the old, may very well turn out to be quite similar to their predecessors. There is no good reason to believe that most linguists have been wrong most of the time. For example, in insisting that word-classes have to be completely rethought in the light of corpus evidence of the similarity of words in their corpus patterns, corpus-driven linguists are not denying the "existence" of nouns and verbs. It is more than likely for many languages that such major categories will anchor a future description. But, unlike today, they will be defined in terms that have scientific rigour, and the members of each class will be enumerable, and the overlap between them will be precisely distinguished.

MULTILINGUAL CORPORA

Translated, or "parallel" corpora have been fashionable for several years now, despite some misgivings people have had about the possible distortion of the text in the target language. Their popularity grew because of what appeared to be early successes in designing automatic alignment software. As long as translations preserve boundaries that can be easily recognised by computers, such as paragraph and sentence breaks, several aligners can produce fair results, though often with a proportion of manual preprocessing. In realistic applications such as the TELRI one, recently reported, (ref) the existing software is of very limited use because most translations are not very tidy with respect to the original.

Nevertheless, the prospects of automatic alignment at phrase level (which is when the interest of human translators is aroused) using a combination of external and internal (linguistic) criteria, are quite good (ref). If so we may anticipate that parallel corpora will be used as sources for future lexicography. As repositories of translators' knowledge, they constitute an important resource, and there is a very large amount of suitable corpus material being published in multilingual communities all over the world. So much so that, as with normal corpora, they have the potential of getting far too big for any manual processing to be involved. The size will be of great value in picking up nuances of translation, but will require fully automatic software to be developed.

It must however be remembered that such corpora do not constitute generic resources in the normal meaning of the term, in that they are not representative samples of the languages concerned. Inevitably they arise only in certain areas of human behaviour, the more formal political, technical, commercial and legal sides principally, with some journalism; so a general corpus of a language, including informal. impromptu, spoken material and so on will not be possible to build without a large translation effort specific for the corpus. Also, it is widely held (ref) that translated language has characteristics that are to do with the translation and not the language, and therefore a corpus of translated text is not a good model of the language as a whole.

We can therefore anticipate a shift of emphasis to "comparable corpora", heralded by the EU PAROLE project, where the texts in each language are chosen through a comparison of the linguistic behaviour of the communities concerned. It has become clear that some of the normal criteria for corpus categorisation are culturally specific, and so the research suggested above on external criteria looks to be needed in multilingual work as much as in monolingual.

SPOKEN LANGUAGE

Spoken and written corpora should be developed together; this is generally agreed. Because of the traditions of scholarship in many countries, the study of phonetics has become

somewhat specialised, and requires its own data and means of data acquisition. A corpus of spoken language can thus be two rather different things — a bank of recorded data for phonetics or "speech" research, or a bank of transcribed spoken data for general linguistic research. There are many intermediate possibilities, of course, since speech data may be transcribed in various ways, and the conventions for spoken material in a general corpus are not as rigid as for the written language.

The two traditions work side-by-side; as each develops opportunities for sharing corpora. Within the European Community it has been decided to separate the development of written and spoken corpus resources, entrusting responsibility for spoken data to the speech scientists. This decision was probably taken on economic grounds, and is questionable professionally; but the danger is that consideration of the spoken language in large reference corpora is given a very low priority. This clashes with the strong opinion expressed by corpus users that the spoken elements in corpora are particularly important.

The key to the harmonisation of collections of speech and writing is the orthography; whatever the linguistic analysis, normal spelling and punctuation should always be retained. If, as is often the case, further sophistications of annotation are provided, they should always be placed carefully relative to the running text and not obscure it.

The normal orthography of the written form of a language has a property that is often overlooked by highly trained specialists; it is legibility. People can read it effortlessly. Everyone

working in the field of language text resources should make a commitment to retain this quality no matter how convenient it may be for computational purposes to make alterations to the orthographic record.

It has also been pointed out that secretarial training makes available large numbers of people able to transcribe a spoken text into normal orthography, and this enables the high cost of building a corpus of spoken language to be kept within bounds. Any interference with the conventions, even apparently trivial additions like noting pauses, are found to slow down transcription by huge factors. The typist is asked to be a covert phonetician, and finds that such decisions form a major distraction from the routine of secretarial transcription.

The emergence of digitised recordings, and of automatic tools for analysis of the sound wave, makes it increasingly possible that further correlations between speech and writing will emerge, and this is a timely research line to encourage, because of its usefulness in applications. The mutual predictability of punctuation and prosodics is particularly important and promising, and success there would make a substantial contribution to speech recognition.

LEXICONS

In this area is the greatest likelihood of considerable overhaul in the coming years. The theories of language used to design

present-day machine lexicons are well-suited to capturing the salient features of scientific terminology, where meaning is particularly clear-cut and relatively invariable. Much of the supporting argument for lexicons of this type concerns information retrieval in relation to technical literature, and is based on assumptions about the stability of the message content in communication (ref).

But these lexicons are quite unable to handle the ordinary patterns of vocabulary. When applied to the lexis of a language they are expensive to construct and poor in performance, because the model is inappropriate for the data — a good example of corpus-based linguistics failing to be sensitive to the evidence of corpora.

New models for the lexicon are essential in the coming decade, to cope with the information about collocation, grammatical co-selection, semantic harmonisation and attitudinal meaning that are among the first types of structural patterning set out in corpus research. These have very little to do with lexicons based on genus-species classification, and abstract semantic features; but they arise from the study of usage and deal with meaning, and so have a clear relevance to the applications of language engineering.

Because of the large investment in lexicons built on pre-corpus models, we can expect considerable resistance to change in this area, especially since the adjustments required are fairly fundamental. They include the following:

~ the features of the new lexicon will be lexicogrammatical, not making a prior division into syntax and lexis, but seeing syntactic statements as more abstract expression of lexical ones.

~ corpus work shows that the meaning of a word or phrase has a profound effect on the choices of other words and phrases in its environment; this interdependence has never been formalised and yet is clearly fundamental, and will have to be incorporated into lexicon design

~ it is inefficient to regard the word as the unit of meaning. At present most lexicons, like dictionaries, arrange meaning around the word; where absolutely unavoidable, phrases are cited as separate meaning-carriers. The new lexicons are likely to be based on units of several words, with a considerable amount of internal variation.

~ in summary, the full force of linked choices on the syntagmatic axis will be incorporated, in contrast to current models, which recognise the importance of syntagmatic choices such as valency, but keep them within an impermeable section of the grammar.

GENERAL

This exercise in summarising trends has led to several predictions of forthcoming changes. They all share a feature which may inhibit their easy acceptance by the community of work-

ers in language text analysis; that is that the new developments rest on models of language that are fundamentally simpler than the old. Here lies the greatest obstacle to acceptance. It is extremely difficult to persuade committed scholars that much of the complex and sophisticated apparatus that has grown up around a subject is not necessary, and people will fear for their reputations and their future work.

Because what is emerging from this are the following messages:

- texts in corpora are best left unannotated
- tools should work on such "raw" text or on the output of other tools
- there should no longer be manual intervention in the preparation of corpora or the application of tools
- current models for lexicons will be confined to the area of terminology

Although all these changes can be understood as inevitable given the nature of corpora, they add up to a formidable problem of conceptualisation.

The reason that such a situation has arisen in corpus linguistics is that it has not yet formulated its own theories and descriptive methods. After three decades of neglect by mainstream linguistics, corpora began to be seen as a useful resource by some computational linguists who had developed elaborate descriptive procedures but had found them incapable of handling ordinary texts. The theories and descriptions were imported

uncritically, and great efforts were then made to use information gleaned from corpora to improve the functioning of the procedures. This is one of the main varieties of corpus-based linguistics.

It is now becoming apparent that the problem is more fundamental than merely adjustment of existing methods. The original theoretical position lying behind these models was antipathetic to data. It rejected usage as being of any interest in as evidence for description, and insisted that the grammars to be produced would be tested by introspection and not by external evidence. Actual speech and writing were specifically excluded from attention in the basic design of the theories. It is not surprising, then, that the descriptive problems encountered seem at times insurmountable, and that few accept the likelihood of simple solutions.

A number of scholars who are not committed to the mentalist ideology of the majority have taken a middle course, and tried to replicate the so-called "consensus" grammar. This is a kind of description well suited to human beings who are fluent in the language concerned, but it cannot easily be converted to the automatic description of corpora. The practical problems are not very different from those of the mentalists - essentially the descriptions do not fit the facts in two ways: in the first instance they are not precise enough, and in the second they ignore vital areas of patterning that if incorporated would improve the precision remarkably. This is the other main kind of corpus-based linguistics.

As a problem of reconceptualisation, this is a serious matter, and those who consider themselves to be shapers of language resources for the future bear a grave responsibility. As the millennium approaches, and a time for change and regrowth, it is a great advantage that the problem has been set out clearly. Those advocating that we should prepare for change, and make plenty of room for it, even though the actual permanent change may not be enormous (we just don't know), feel that the potential for understanding language with the aid of computers is very great, and should not be hampered by methodological conventions dating back before corpora were available.



附录 1

主 要 术 语

语料库（corpus；或 corpora，corpuses[复]）

语料库是指按照一定的语言学原则，运用随机抽样方法，收集自然出现的连续的语言运用文本或话语片段而建成的具有一定容量的大型电子文本库。从其本质上来说，语料库实际上是通过对自然语言运用的随机抽样，以一定大小的语言样本代表某一研究中所确定的语言运用总体。语言学者对语料库的定义不尽相同，辛克莱(Sinclair 1996)认为语料库是"按照明确的语言学标准选择并排序的语言运用材料汇集，旨在用作语言的样本"。阿特金斯等(Atkins & Clear1992)认为语料库是"按照明确的设计标准，为某一具体目的而集成的大型文本库"。赫努(Renouf 1987:1)认为，语料库是"由大量收集的书面语或口头语构成，并通过计算机储存和处理，用于语言学研究的文本库"。

学习者语料库（computer learner corpus）

学习者语料库是通过收集语言学习者各种书面语和口语的自然语料而建立的一种学习者语言数据库。现代计算机学习者语料库（简称CLC，参见 Granger 1996）与以往的学习者错误语料库不同，其目的在于对学习者的语言特征和语言发展进行全面和系统的对比分析，而不仅限于对学习者错误进行分析。学习者语料库的优势在于能够提供有关学习者语言发展的全面信息。在外语学习和教学中，基于学习者语

料库的研究能提供学习者典型困难以及在某一具体方面的主要障碍的反馈信息。另外,通过不同类型学习者语言的对比,可以发现学习者在某一发展阶段的共同语言特征和个体特征,从而促进外语教学在大纲设计、教材开发以及课堂教学诸方面更加适应学习者的需求。

词语索引 (concordance)

词语索引的基本意义是把搜索词或词组按字母或频率顺序排列与其所在语境一同展示。词语索引最常见的形式是 KWIC,即"语境中的关键词"。KWIC 技术为每一个搜索到的关键词提供所在行固定数量的语境词,并以该关键词为中心在屏幕上显示出来。词语索引软件在文学和语言学分析中的应用包括搭配分析、主题分析、用于词典编纂的针对某一词汇的例句援引、语音分析、词素分析、语义及话语分析等。

词频统计 (word frequency count)

在对语料库文本进行统计分析中,词频统计和文本统计功能把语料库中出现的所有"类符"(type) 统计列表。通常,词表提供以下几种信息:1)类符总数。词表中每个类符按字母顺序和频数顺序排列,可以进行顺序和逆序排列;2)每个类符的频数。类符的频数即该类符的形符(token) 总量。通过观察形符的分布,可以得到某一文类文本中词汇的基本分布信息,并通过编辑,提取常用词汇;3)每个类符的频率。

关键词 (key word)

关键词是语料库用户在对某一词形或用法在语料库中索引时所键入的词形、词组、或带有通配符的字母组合。用户通过关键词对词语索引进行控制,以获得所期望的数据。某些词语索引分析软件也将关键词称作搜索词 (search word, 简称 SW)、节点词 (node word)等。以关键词为中心左右显示的词数构成了该词的"跨距"(word span)。"跨距"中的词构成了搜索词的微型语境,或"同现语篇"(co-text)。在对某一关键词进行检索时,可通过键入语境词作为限制条件。语境词可以是与关键词具有搭配关系的词或词组。在检索中可以围绕任何关键词,从该词所在行、段

落以至整篇文本无限扩充显示。这种方法对单个的关键词不仅提供短语和句子层面上的使用语境，还可提供整个语篇语境。

型式统计 (pattern)

词语型式统计即根据索引中所规定的跨距，计算在跨距内每个词位词语出现的频数，并按频数降序排序。词语型式统计直观展示某一搜索词前后各个词位的词语分布情况。根据词语型式统计，可以观察词语搭配关系以及围绕某一搜索词不同的聚集词群。

主题词 (theme word)、词图 (plot)

在不同的文本中，词语的选择和分布存在差异。不同文本的主题通过词语的选择体现出来。通过提取和分析文本中具有超常频率的词以及具有共现关系的词语或词群，可以确定文本的主题和表达该主题的词集，进而研究作者对某一主题的心理词符和知识表达问题。就学习者语料库而言，通过主题词研究，可以观察学习者对某一主题所使用的词语以及词语之间的关系。主题词提取的方法首先是对比某一完整连续文本和一个更大的参照语料库(reference corpus)，把文本中差异显著的词语提取出来，生成一个主题词表。差异的显著性通过 X^2 检验用 P 值表示($P \leqslant 0.00001$)。词图 (plot) 统计是根据主题词表，计算出各个主题词在文本中的位置分布。词图的意义主要在于对某一连续文本的词语分布进行统计和计算。如果是像 JDEST，BROWN，LOB，或 CLEC 这样的杂合性语料，尽管也能统计出单个词的词图，但由于缺乏相对应的文本信息作为参照，其价值有限。另外，单个词图只有同其他词图放在一起进行比较，才能显示出真正意义。对搜索词进行各例词图统计的前提必须是所检索的文本是一连续的整体，如一篇小说，或论文。通过词图统计，可以观察主题词在文本中出现的先后顺序和分布密度，直观地分析该文本的主题发展或情节推进与词语的关系。

词语学 (lexis)

词语学是研究词汇行为的科学。它以真实语言使用中的实例为研究材

料，以词形(word form)在文段中的分布、位置和观察频数等为基本数据来研究语言的词语结构，即词形是如何组织成为搭配和成语等语言单位的。词语学不同于语义学，后者主要研究语言的意义体系，而前者将词形的搭配型式与意义差别结合一体研究。词语学也不同于语法，后者强调语法规则的一般概括作用，而前者强调词的个体行为型式。词语学是弗斯(Firth)语言学研究的核心内容之一。

验证数据 (attested data)

验证数据又叫自然发生数据(naturally occurring data)，与直觉数据(intuitive data)、诱导数据(elicitative data)相对。验证数据指的是在真实语言使用中出现的证据。使用验证数据是实证主义研究的重要特点之一，代表了一种学术思想和研究方法。在实证主义研究者看来，语言学研究必须采用验证数据，用数据驱动的方法(data-driven approach)才能有效地发现语言运作的机制、规律和特点，而语言学家的直觉和语言学研究活动中的诱导数据都很难完全反映真实语言的使用与运作。在应用语言学和语言教学研究中，验证数据常被称之为真实数据(authentic data)。

类联接 (colligation)

类联接指文本中有关语法范畴的结合。类联接不是与词语搭配平行的抽象，而是高一级的抽象。类联接是关于句法结构和范畴的表述，搭配则是类联接在词语层面的具体实现。严格意义上的词语搭配应当是指位于同一语法结构内的词与词之间的组合。因此，类联接就是词语搭配出现于其中的语法结构和框架。一个类联接代表了一个类别的词语搭配，可称为搭配类(collocational class, collocational type)。比如，N + V 就是一个类联接，它代表一类搭配，在研究中常称之为 N + V 搭配。类联接和词语搭配是紧密相联的；在研究中，一个类联接的确立必须以足够的词语搭配行为为依据，词语搭配行为的研究则应在类联接的框架内进行。

词语搭配（collocation）

词语搭配是在文本中为实现一定的意义从而以一定的语法形式因循组合使用的一个词语序列，构成该序列的词语相互预期，以大于偶然的机率共现。该定义较为明确地体现了词语搭配的界定特点：

1. 词语搭配是组合轴上融意义与形式于一体的词汇共现现象。
2. 词语搭配是因循性的语言使用。
3. 词语搭配是文段中反复重现的词汇结合方式。
4. 词语搭配在很大程度上受语域的影响或制约。
5. 词语搭配是一个词语序列的重现。
6. 词汇间相互搭配的关系可由统计度量确定。

有些研究者不主张考虑语法对词语行为的限制作用，所以他们关于词语搭配的定义一般都只着重强调"词语的共现"这一条件和标准。

节点词（node word）

节点词即词语搭配研究要调查其行为的关键词(key word)。人们每一次在语料库中查询，都要将自己要研究的某个词从键盘上输入计算机，计算机则按照编好的程序，用词语索引(key word in context)的方法显示出词语索引。在每一行词语索引中，关键词，也就是节点词，总是居中出现，而左右则是构成其语境的词语，研究者可据此分析其行为。语料库中的每个词都可以是节点词。选取哪些词为节点词完全由研究者根据其研究内容和研究目的而定。

跨距（span）

跨距指的是节点词左右的语境，以词为单位计算，不包括标点符号。假如将跨距定为－4/＋4，意思是说在节点词左右各取4个词为其语境。

距位（span position）

距位说的是跨距内每一个词所处的相对位置，常用N－1，N＋1，N－2，N＋2等表示。有些研究者用Left1，Left2，Right1，Right2等来表示。

共现词 (co-occurring words)

所有落入限定跨距内的词都将被视作节点词的共现词。这些共现词是研究和描述节点词行为的依据。

偶然共现词 (casual collocates)

那些落入界定跨距内但对节点词没有一般预见作用的词语在研究中被称为偶然共现词。严格说来,偶然共现词对真正意义上的词语搭配研究没有太大意义,应当排除。排除的办法之一就是通过统计检验。一般来说,这类词重现的概率较低。当一个词在语料库中出现的频数较高时,统计检验会突出它的典型搭配;它的偶然共现不太可能具有统计意义上的显著性而被过滤掉。再之,通过类联接即语法框架的建立,也可将偶然共现词排除。

非连续搭配 (discontinuous collocation)

非连续搭配又被称为非邻近搭配(discontiguous collocation)。有些常见的词语搭配往往是习惯性地被分隔而处于非连续的位置上。比如"different from"这个搭配,如果它以下列形式出现在文本中,那就是非连续搭配。例如:Non-cooperators were not <u>different</u> in age or other conventional factors <u>from</u> the rest.

显著搭配 (significant collocation)

显著搭配是指那些构成成分共现次数比人们根据它们在文本中各自出现的频数做出的预测还要多的词语序列。词语搭配行为从根本上来说是个概率问题。在词语行为研究中,研究者通常用概率的方式来描述搭配的典型程度。这就要对搭配进行统计检验,在统计上达到一定显著标准的搭配就在一定程度上表明它的典型性。

话语社团 (discourse community)

话语社团不同于语言社团,它和人们所从事的专业活动密切相关。一个话语社团有大体上一致和公开的共同目标,比如计算机科学家和语

言学家有各自具体的研究目标；有各自供社团成员交流信息的机制，比如专业研究会、协会、学会等组织，有各自的研讨会、信息通报、通讯往来等活动；有各自为促进目标实现而用于思想交流的文类形式或语篇发展形式，以制约人们选取适当的专题、采用得体的形式、完成一定的功能以及决定语篇要素的位置等，比如各学科领域办的期刊、学报等所规定的论文的形式与结构；它还有各自专用的词语系统，比如技术词汇、次技术词及其特有的用法。因此，话语社团内部有一套高度因循性的交流方式，包括文类形式、语篇模式和词语的选择与组合。这些方式是话语社团特有的，外行人员通常对之不甚了解。

搭配范围（collocational range）

节点词的搭配力使其常常和一个系列的词形形成搭配伙伴，结合一起使用。这个系列的词就构成了节点词的搭配范围。如在学术英语中，*experiment* 的搭配范围大致包括 *performed*、*conducted*、*carried out*、*presented*、*described*、*controlled*、*demonstrates*、*shows*、*indicates*、*suggests*、*further*、*laboratory*、*preliminary*、*present*、*proposed*、*simulation*、*results*、*data* 等词形。搭配范围这一术语由麦金托什(A. McIntosh)提出。韩礼德和辛克莱等学者则用词语型式(lexical pattern)等术语。

期望频数（expected frequency）

在词语搭配研究中，期望频数是个很重要的概念。如果把一个语料库包含的总词数假定为 W，而某个搭配词与节点词共现的频数为 C，2S 为限定的跨距，N 为节点词在语料库中出现的频数，那么这个搭配词与节点词共现的期望频数就是：

$$Ef = C * (2S + 1) * N / W$$

一个搭配序列在语料库中实际出现的频数越高于其期望频数，它就越显著。期望频数被用于搭配词的显著性统计度量中，如 Z 分值(或 T 分值)检验。

词丛 (cluster)

词丛指的是由两个或两个以上词形构筑的连续的序列。按所含词形的多少,词丛可以分为2词词丛、3词词丛、4词词丛等。词丛可以通过一些语料库分析软件自动提取,并可以自动计算出现频数,其原理和单个词的词频计算相似。词丛所反映的是一种比搭配词更紧密的关系,更像词组,但又不同于一般意义上的词组,因为有很多词丛不是词组。任何词都有与其他词成群出现的趋势。词丛对类联接、搭配以及语义韵的研究有重要的参考价值。

语料档案 (archive)

一些语料库语言学家使用该术语,把它与"语料库"区分开来。与"语料库"不同的是,语料档案没有明确的收集标准,即不限定"文类"(genre),更不规定各类的比例,因此其文本的收集相对要随意得多。"语料库"有确定的规模,而语料档案则不一定,它可以是开放的。按照这种区分,BROWN、LOB是"语料库",而Bank of English就应该是语料档案。

词类赋码 (tagging)

所谓词类赋码就是对文本中每一个词形(包括标点符号)赋予相应的词类码。词类码一般是按词的语法特征来分类的,所以也称作语法码。词类赋码是自然语言处理中最基本的一步。目前,自动赋码系统有很多,各种赋码系统所用的词类码的数量和类型都不尽相同,所依据的赋码原理也不尽相同。许多自动词类赋码系统利用了语言中的概率信息,赋码准确率达到96%以上。词类赋码不仅用于自然语言处理,也可用于语言研究,附加词类码的语料库能为某些语言研究提供极大便利。

句法分析 (parsing)

所谓句法分析就是对文本中的每一个句子进行句法标注。句法分析一般要建立在词类赋码的基础上,需要利用语言中的概率信息。相对词

类赋码系统而言，句法分析系统的数量要少得多，准确率也远远没有词类赋码准确率高。

形符 (tokens)

形符指两个相邻空格之间的连续字符串，一个文本的形符数就是该文本的长度。文本一共有多少个词，就有多少个形符。例如，某文本的长度是 1000 个词，那么它的形符就是 1000 个。在词频统计中单纯的形符意义并不大，其意义主要体现在“标准化类符形符比”之中。

类符 (types)

类符在词频统计中相同的形符称为类符，一个文本的类符数就是该文本不同词形的数量。文本一共有多少个不同的词形，就有多少个类符。例如，某长度为 1000 词的文本里不同的词形是 500 个，那么它的类符就是 500 个。在词频统计中单纯的类符意义也并不大，其意义主要也体现在“标准化类符形符比”之中。

标准化类符/形符比 (standardized type/token ratio)

标准化类符/形符比是为弥补未经标准化的类符/形符比之不足而使用的一个统计量。它的计算方法是按一定的长度分批计算文本的类符/形符比，然后求出它们的平均值。这个长度可以根据需要自行设定，假设定为 1000 个词，则先计算第一个 1000 词的类符/形符比，然后计算第二个 1000 词的类符/形符比，如此计算下去一直到文本的结束，(最后所剩未满 1000 词的语料就不予考虑，)最后计算它们的平均值就得到标准化类符/形符比。这样计算出来的比率就可以用来比较不同长度文本的用词变化，弥补了未经标准化的类符/形符比的缺陷。

词频 (word frequency)

词频指各词形(word form)在文本中出现的频数，是语料库分析中最基本的统计量之一，可以通过分析软件自动提取。词频统计结果可以直接用于观察，比较词的使用频率，还可以进一步用于其他更为复杂的

统计之中，如“搭配力”、“关键性”的计算之中。

搭配力（collocability）

搭配力指节点词或搜索词对搭配词的吸引力，一般用 Z 分数（Z-score）或 T 分数（T-score）来表示。搭配力越大，节点词对搭配词的吸引力越大，它们之间搭配的意义也就越大；反之，搭配力越小，节点词对搭配词的吸引力越小，两者间搭配的意义也就越小。在计算搭配力之前必须知道 5 个量，它们分别是：跨距、节点词在整个文本中的频数、搭配词在整个文本中的频数、搭配词与节点词共现的频数以及整个文本的长度。



附录 2

书面英语词类码

说明: 下表共分四列,第一列为词类码(其中左侧为 BROWN 语料库中的赋码,右侧其他语料库中的变异情况);第二列为该词类码所用于的语料库,其中 ALL 代表用于以上所有语料库;第三列为词类码的名称,即该词类码所代表的意义;第四列为用斜体表示的例证(部分并入第三列)。此外,在第三、第四列中若内容与上一行相同,则以 ibid 代替。下表中的略语及其代表的语料库如下: Br = BROWN; Go = Gothenburg; La = Lancaster; Ls = Leeds; LOB = LOB; Os = Oslo。

Tags		Used in	Meaning	Examples
.		Br, Go, Ls, Os	sentence closer	. ; ? !
	.	LOB, La	full stop	
	;	LOB, La	semi-colon	
	?	LOB, La	question mark	
	!	LOB, La	exclamation mark	
(		ALL	left bracket (round or square)	

（续表）

Tags		Used in	Meaning	Examples
)		ALL	right bracket (round or square)	
,		ALL	comma	
:		ALL	colon	
—		ALL	dash	
	*-	LOB	ibid	
	-	La	ibid	
	...	LOB, La	ellipsis	
	&FO	LOB, La	formula	
	*'	LOB	open inverted commas	
	**'	LOB	close inverted commas	
	"	La	open and close inverted commas	
	$	La	genitive inflexion > 's >'	
*		Br, Ls, Os	*not* > *n't*	
	XNOT	LOB	ibid	
	G	Go	ibid	
	XX	La	ibid	
ABL		ALL	pre-qualifier	*quite, rather*
ABN		Br, Go, Ls, LOB, Os	pre-quantifier	*half, all*
	DB	La	ibid	ibid

（续表）

Tags		Used in	Meaning	Examples
ABX		Br, Go, Ls, LOB, Os	pre-quantifier and double conjunction	*both*
	LE	La	leading co-ordinator	*both* before *and*
	DB2	La	plural before-determiner	*both* without *and*
AP		Br, La, Ls, LOB, Os	post-determiner	*many, much, next, former, other*
	Q	Go	ibid	ibid
	ATA	La	after-article	*other, only*
	DA	La	after-determiner capable of pronominal function and neutral for number	*such, former, same*
	DA1	La	singular after-determiner	*little, much*
	DA2	La	plural after-determiner	*few, several, many*
	DA2R	La	comparative plural after-determiner	*fewer*
	DA2T	La	superlative plural after-determiner	*fewest*
	DAR	La	comparative neutral after-determiner	*more, less*
	DAT	La	superlative neutral after-determiner	*most, least*

（续表）

Tags		Used in	Meaning	Examples
AP $		ALL	possessive post-determiner	*other's*
AT		Br, Ls, Os	article	*the, a, every, no*
	T	Go	definite article	*the*
	F	Go	indefinite article	*a*
	Q	Go	quantifier	*every*
	AT	LOB	singular article	*a, every*
	AT1	La	ibid	ibid
	ATI	LOB	neutral article	*the, no*
	AT	La	ibid	ibid
BE		Br, Ls, LOB, Os	*be*	
	VBO	La	ibid	
	B	Go	indicative *be*	
	BM	Go	imperative *be*	
	BB	Go	subjunctive *be*	
BED		Br, Ls, LOB, Os	*were*	
	VBDR	La	ibid	
	BD	Go	indicative *were*	
	BDB	Go	subjunctive *were*	
BEDZ		Br, Ls, LOB, Os	*was*	
	BD	Go	*was*	
	VBDZ	La	ibid	

（续表）

Tags		Used in	Meaning	Examples
BEG		Br, Ls, LOB, Os	*being*	
	BG	Go	ibid	
	VBG	La	ibid	
BEM		Br, Ls, LOB, Os	*am*, > *'m*	
	B	Go	ibid	
	VBM	La	ibid	
BEN		Br, Ls, LOB, Os	*been*	
	BN	Go	ibid	
	VBN	La	ibid	
BER		Br, Ls, LOB, Os	*are*, > *'re*	
	B	Go	ibid	
	VBR	La	ibid	
BEZ		Br, Ls, LOB, Os	*is*, > *'s*	
	B	Go	ibid	
	VBZ	La	ibid	
CC		Br, Ls, LOB, Os	co-ordinating conjunction	*and*, *but*, *or*, *yet*
	Y	Go	ibid	ibid
	CC	La	general co-ordinating conjunction	

（续表）

Tags		Used in	Meaning	Examples
	CCB	La	*but*	
	CF	La	semi-co-ordinating conjunction	*so, then, yet*
CD		Br, Ls, Os	cardinal numeral	*two, 6, hundred, 2.34*
	L	Go	ibid	ibid
	CD	LOB	general cardinal	
	MC	La	cardinal	*two, 6, 2.34*
	NNO	La	numeral noun, neutral for number agreement	*dozen, hundred*
	CD1	LOB	*one, 1*	
	MC1	La	ibid	
	CD-CD	LOB	hyphenated number	*42 – 3, 1770 – 1827*
	MC-MC	La	ibid	ibid
CS		Br, LOB, Os	subordinating conjunction	*if, although, as, because*
	Z	Go	ibid	ibid
	CS	Ls	subordinating conjunction other than *as*	
	CS	La	general subordinating conjunction	
	CSH	La	*the* introducing comparative clauses	

（续表）

Tags		Used in	Meaning	Examples
	CSN	La	*than* as conjunction	
	CST	La	*that* as conjunction	
	CSW	La	*whether* as conjunction	
	CSA	Ls, La	*as* as conjunction	
DO		Br, Ls, LOB, Os	*do*	
	VDO	La	ibid	
	D	Go	indicative *do*	
	DM	Go	imperative *do*	
DOD		Br, Ls, LOB, Os	*did*	
	DD	Go	ibid	
	VDD	La	ibid	
DOZ		Br, Ls, LOB, Os	*does*	
	D	Go	auxiliary *does*	
	VDZ	La	*does*	
DT		Br, Ls, LOB, Os	singular determiner	*this, that, another*
	E	Go	ibid	ibid
	DD1	La	ibid	ibid

（续表）

Tags		Used in	Meaning	Examples
DT $		ALL	possessive singular determiner	*another's*
DTI		Br, Ls, LOB, Os	neutral determiner capable of pronominal function	*any, some*
	Q	Go	ibid	ibid
	DD	La	ibid	ibid
DTS		Br, Ls, LOB, Os	plural determiner	*these, those*
	ES	Go	ibid	ibid
	DD2	La	ibid	ibid
DTX		Br, Go, Ls, LOB, Os	determiner and double conjunction	*either, neither*
	LE	La	leading co-ordinator	*either* before *or*
	DD1	La	Singular determiner	*either* without *or*
EX		Br, Ls, LOB, Os	existential *there*	
	X	Go	ibid	
	EX	La	ibid	
FW-		Br, Go, Ls, Os	foreign word	
	&FW	LOB, La	ibid	
	HH	Ls	pre-infinitive	*so as, in order*

（续表）

Tags		Used in	Meaning	Examples
	BCS	La	before-conjunction	*in order, even* before *that/if*
	BTO	La	before-infinitive	*in order, so as* before *to*
-HL		Br	word in headline	
HV		Br, Ls, LOB, Os	*have*	
	H	Go	ibid	
	VHO	La	ibid	
HVD		Br, Ls, LOB, Os	*had*, >*'d* (preterite)	
	HD	Go	ibid	
	VHD	La	ibid	
HVG		Br, Ls, LOB, Os	*having*	
	HG	Go	ibid	
	VHG	La	ibid	
HVN		Br, Ls, LOB, Os	*had* (past participle)	
	HN	Go	ibid	
	VHN	La	ibid	
HVZ		Br, Ls, LOB, Os	*has*, >*'s*	
	H	Go	auxiliary *has*, >*'s*	
	VHZ	La	*has*, >*'s*	

（续表）

Tags		Used in	Meaning	Examples
IN		Br, LOB, Os	preposition	
	P	Go	ibid	
	IN	Ls	general preposition	
	II	La	ibid	
	ICS	La	preposition-conjunction of time	*after*, *before*, *since*, *until*
	INF	Ls	*for* as preposition	
	IF	La	ibid	
	INO	Ls	*of*	
	IO	La	ibid	
	INW	Ls	*with*, *without*	
	IW	La	ibid	
JJ		Br, Ls, Os	adjective	
	J	Go	ibid	
	JJ	LOB, La	general adjective	
	JA	La	predicative adjective	*tantamount*, *afraid*, *asleep*
	JK	La	adjective catenative	*able* in *be able to*, *willing* in *be willing to*

（续表）

Tags		Used in	Meaning	Examples
	JJB	LOB	attributive adjective	*utter, late, model* in *a model prisoner, 15-nation* in *the 15-nation NATO council*
	JB	La	ibid	ibid
	JNP	LOB	adjective with w. i. c.	*Welsh, Keynesian*
JJS		Br, Go, Ls, Os	semantically superlative adjective	*main, chief, top, uppermost*
	JJB	LOB	ibid	ibid
	JB	La	attributive adjective	*main, chief, top*
	JBT	La	attributive superlative adjective	*utmost, topmost*
JJR		Br, Ls, LOB, Os	comparative adjective	*older, better, stronger*
	JR	Go	ibid	ibid
	JJR	La	general comparative adjective	*older, better, stronger*
	JBR	La	attributive comparative adjective	*upper, outer*
JJT		Br, Ls, LOB, Os	morphologically superlative adjective	*oldest, best, strongest*
	JT	Go	ibid	ibid
	JJT	La	ibid	ibid

（续表）

Tags		Used in	Meaning	Examples
MD		Br, Ls, LOB, Os	modal auxiliary	*can, may, would, ought*
	M	Go	present modal auxiliary	*can, may*
	MD	Go	preterite modal auxiliary	*might, would*
	VM	La	modal auxiliary	*can, may, would*
	VMK	La	modal catenative	*ought, used*
-NC		Br, Go, Ls, Os	cited word	
	NC	LOB	ibid	
	NC2	La	plural cited words	as in *too many ifs and buts*
NN		Br, Ls, Os	singular common noun	*book, girl, sheep, headquarters* governing *is*
	N	Go	ibid	ibid
	NN	LOB	singular common noun	*book, girl, sheep*
	NNP	LOB	singular common noun with w.i.c.	*Londoner, Luddite*
	NN	La	common noun neutral for number	*sheep, cod, headquarters*
	NN1	La	singular common noun	*book, girl*

(续表)

Tags		Used in	Meaning	Examples
	NNJ	La	organization noun neutral for number	*Company, Co., group*
	NNT1	La	singular temporal noun	*day, week, year*
	NNU1	La	singular unit of measurement	*inch, kilo*
	NNU	LOB, La	abbreviated unit of measurement neutral of number	*in., kg*
	NPL	LOB	singular locative noun with w.i.c.	*Island, Street*
	NNL	La	locative noun neutral for number	*Is.*
	NNL1	La	singular locative noun	*island, Street*
	NPT	LOB	singular titular noun with w.i.c.	*Bishop, Mrs*
	NNS1	La	singular titular noun	*Mrs, President, Rev.*
	NNSA1	La	following abbrev. singular titular noun	*M.A.*
	NNSB	La	preceding abbrev. titular noun	*Rt. Hon.*
	NNSB1	La	preceding abbrev. singular titular noun	*Prof.*
	ZZ	LOB	letter of the alphabet	as in *it fitted to a T, if x is greater than y*

（续表）

Tags		Used in	Meaning	Examples
	ZZ1	La	singular letter of the alphabet	
	ZZ2	La	plural letter of the alphabet	
NN $		Br, Ls, Os	possessive singular noun	*girl's*
	NX	Go	ibid	ibid
	NN $	LOB	general possessive singular noun	
	NNP $	LOB	possessive singular common noun with w.i.c.	*Institution's*
	CD $	LOB	possessive cardinal	*two's complement*
	CD1 $	LOB	*one's*	
	NNU $	LOB	possessive abbreviated neutral unit of measurement	*cwt's*
	NPL $	LOB	possessive singular locative noun with w.i.c.	*Island's*
	NPT $	LOB	possessive singular titular noun with w.i.c.	*Mayor's*
	OD $	LOB	possessive ordinal	*sixth's*

（续表）

Tags		Used in	Meaning	Examples
	NN1 $		singular possessive nouns not containing apostrophe	*domini* in *anno domini*
NNS		Br, Ls, Os	plural common noun	*books, girls, sheep, Associates, headquarters*
	NS	Go	ibid	ibid
	NNS	LOB	plural common noun	*books, girls, sheep*
	NNPS	LOB	plural common noun with w.i.c.	*Englishmen*
	NN2	La	plural common noun	*books, girls*
	NNJ2	La	plural organization noun	*groups, councils, unions*
	NNT2	La	plural temporal noun	*days, weeks, years*
	APS	LOB	plural post-determiner	*others*
	CDS	LOB	plural cardinal	*threes, 3s, hundreds*
	MC2	La	plural cardinal number	*threes, 3s*
	MF	La	fraction neutral for number	*two-thirds*
	NNO2	La	plural numeral noun	*hundreds, millions*

（续表）

Tags		Used in	Meaning	Examples
	CD1S	LOB	*ones*	
	NNUS	LOB	plural unit of measurement	*ins.*, *feet*
	NNU2	La	ibid	ibid
	NNLS	LOB	plural locative noun with w.i.c.	*Islands*
	NNL2	La	ibid	ibid
	NPTS	LOB	plural titular noun with w.i.c.	*Presidents, Messrs*
	NNS2	La	plural titular noun	*Presidents*
	NNSB2	La	preceding abbrev. plural titular noun	*Messrs*
NN-S $		Br, La, Ls, Os	possessive plural noun	*girls', children's*
	NSX	Go	ibid	ibid
	NNS $	LOB	possessive plural noun	
	APS $	LOB	possessive plural post-determiner	*others'*
	NNPS $	LOB	possessive plural common noun with w.i.c.	*Associates'*

（续表）

Tags		Used in	Meaning	Examples
	NNUS $	LOB	possessive abbrev. plural unit of measurement	*c.c.s'*
	NPLS $	LOB	possessive plural locative noun with w.i.c.	*Islands'*
	NPTS $	LOB	possessive plural titular noun with w.i.c.	*Presidents'*
	NRS $	LOB	possessive plural adverbial noun	*Sundays'*
NP		Br, Ls, LOB, Os	singular proper noun	*London, Naismith, October, Andes*
	C	Go	ibid	ibid
	NP	La	proper noun neutral for number	*Indies, Andes*
	NP1	La	singular proper noun	*London, Frederick, Naismith*
	NPM1	La	singular month noun	*October*
NP $		Br, La, Ls, LOB, Os	possessive singular proper noun	*Juliet's*
	CX	Go	ibid	ibid
NPS		Br, Ls, LOB, Os	plural proper noun	*Naismiths, Americas, Octobers*
	CS	Go	ibid	ibid
	NP2	La	plural proper noun	*Naismiths, Americas*

（续表）

Tags		Used in	Meaning	Examples
	NPM2	La	plural month noun	*Octobers*
NP-S $		Br, La, Ls, LOB, Os	possessive plural proper noun	*Naismiths'*
	CSX	Go	ibid	ibid
NR		Br, Go, Ls, LOB, Os	adverbial noun	*home, west, tomorrow, Sunday*
	ND	La	noun of direction	*west*
	NPD1	La	singular weekday noun	*Sunday*
	RT	La	nominal adverb of time	*tomorrow*
NR $		ALL	possessive adverbial noun	*tomorrow's*
NRS		Br, Go, Ls, LOB, Os	plural adverbial noun	*Sundays, yesterdays*
	NPD2	La	plural weekday noun	*Sundays*
OD		Br, Ls, LOB, Os	ordinal numeral	*second, 2nd*
	K	Go	ibid	ibid
	MD	La	ibid	ibid
PN		Br, Ls, LOB, Os	nominal pronoun	*anybody, everyone, nothing, none*
	O	Go	ibid	ibid

（续表）

Tags		Used in	Meaning	Examples
	PN	La	neutral nominal pronoun	*none*
	PN1	La	singular nominal pronoun	*anybody*, *everyone*, *nothing*, *one* as pronoun
PN $		Br, La, Ls, LOB, Os	possessive nominal pronoun	*everybody's*
	QX	Go	ibid	ibid
PP $		Br, Ls, LOB, Os	possessive personal pronoun	*my*, *your*, *our*
	APP $	La	ibid	ibid
	RX	Go	singular possessive personal pronoun	*my*
	RSX	Go	plural possessive personal pronoun	*our*
PP-$ $		Br, Go, Ls, LOB, Os	second possessive personal pronoun	*mine*, *yours*, *ours*
	PP $	La	ibid	ibid
PPL		Br, Ls, LOB, Os	singular reflexive personal pronoun	*myself*, *herself*
	RL	Go	ibid	ibid
	PNX1	La	reflexive indefinite pronoun	*oneself*
	PPX1	La	singular reflexive ersonal pronoun	*yourself*, *itself*

（续表）

Tags		Used in	Meaning	Examples
PPLS		Br, Ls, LOB, Os	plural reflexive personal pronoun	*ourselves, themselves*
	RLS	Go	ibid	ibid
	PPX2	La	ibid	ibid
PPO		Br, La, Ls, LOB, Os	objective personal pronoun	*me, us, him, it, you* in non-subject position
PPS		Br, La, Ls, LOB, Os	3rd singular nominative pronoun	*he, she, it* in subject position
PPSS		Br, La, Ls, LOB, Os	other nominative pronoun	*I, we, they, you* in subject position
	R	Go	*I, me, you, he, him, she, her, it*	
	RS	Go	*we, us, they, them*	
	PP3A	LOB	*he, she*	
	PPHS1	La	ibid	
	PP3	LOB	*it*	
	PPH1	La	ibid	
	PP1O	LOB	*me*	
	PPIO1	La	ibid	
	PP1OS	LOB	*us*	
	PPIO2	La	ibid	
	PP3O	LOB	*him, her*	

（续表）

Tags		Used in	Meaning	Examples
	PPHO1	La	ibid	
	PP3OS	LOB	*them*	
	PPHO2	La	ibid	
	PP2	LOB	*you*	
	PPY	La	ibid	
	PP1A	LOB	*I*	
	PPIS1	La	ibid	
	PP1AS	LOB	*we*	
	PPIS2	La	ibid	
	PP3AS	LOB	*they*	
	PPHS2	La	ibid	
QL		Br, Go, Ls, LOB, Os	degree adverb	*very, so, too*
	RG	La	ibid	ibid
QLP		Br, Go, Ls, LOB, Os	post-adjectival/adverbial degree adverb	*enough, indeed*
	RGA	La	ibid	ibid
RB		Br, Ls, Os	adverb	*quickly, well, understandably*, etc., *else, along* as adverb
	A	Go	ibid	ibid
	RB	LOB	general adverb	
	RR	La	ibid	
	RA	La	adverb after nominal head	*else, galore*

（续表）

Tags		Used in	Meaning	Examples
	REX	La	apposition-introducer	*namely, e.g.*
	RL	La	locative adverb	*alongside, forward*
	RI	LOB	adverb, homograph of preposition	*before, along, by*
RB$		ALL	adverb with possessive suffix	*else's*
RBR		Br, Ls, LOB, Os	comparative adverb	*better, longer, more*
	AR	Go	ibid	ibid
	RGR	La	comparative degree adverb	*more, less*
	RRR	La	comparative general adverb	*better, longer*
RBT		Br, Ls, LOB, Os	superlative adverb	*best, longest, most*
	AT	Go	ibid	ibid
	RGT	La	superlative degree adverb	*most, least*
	RRT	La	superlative general adverb	*best, longest*
RN		Br, Go, Ls, LOB, Os	nominal adverb	*here, there, now, then*
	RL	La	locative adverb	*here, there*
	RT	La	nominal adverb of time	*now, then*

（续表）

Tags		Used in	Meaning	Examples
RP		Br, Go, Ls, LOB, Os	prepositional adverb which is also particle	*across, off, up, about*
	RP	La	prepositional adverb which is also particle	
	RPK	La	prepositional adverb catenative	*about* in *be about to*
-TL		Br	word in title	
TO		Br, Ls, LOB, Os	infinitive *to*	
	U	Go	ibid	
	TO	La	ibid	
UH		Br, La, Ls, LOB, Os	interjection	*hello, no*
	I	Go	ibid	ibid
VB		Br, Ls, LOB, Os	verb, base form	*eat, request*
	VV0	La	ibid	ibid
	V	Go	indicative verb	
	VM	Go	imperative verb	
VBD		Br, Ls, LOB, Os	verb, preterite	*ate, requested*
	VD	Go	ibid	ibid

（续表）

Tags		Used in	Meaning	Examples
	VVD	La	ibid	ibid
VBG		Br, Ls, LOB, Os	present participle	*eating, requesting*
	VG	Go	ibid	ibid
	VVG	La	present participle	
	VVGK	La	present participle catenative	*going* in *be going to*
	VDG	La	*doing*	
VBN		Br, Ls, LOB, Os	past participle	*eaten, requested*
	VN	Go	ibid	ibid
	VVN	La	past participle	
	VVNK	La	past participle catenative	*bound* in *be bound to*
	VDN	La	*done*	
VBZ		Br, Ls, LOB, Os	3rd singular form of verb	*eats, requests*
	V	Go	ibid	ibid
	VVZ	La	ibid	ibid
WDT		Br, Ls, LOB	*wh-* determiner	*what, which, whatsoever, whichever*
	WDT	Os	interrogative *wh-* determiner	*what, whatever,* interrogative *which*

（续表）

Tags		Used in	Meaning	Examples
	WDTR	Os	relative *wh-* determiner	*which*
	DDQ	La	*wh-* determiner without *-ever*	*what, which*
	DDQV	La	*wh-*ever determiner	*whatsoever, whichever*
WP $		Br, Ls, LOB	possessive *wh-* pronoun	*whose*
	WP $	Os	interrogative *whose*	
	WP $ R	Os	relative *whose*	
	DDQ $	La	possessive *wh-* determiner	*whose*
	PNQV $	La	*whosever*	
WPO		Br, La, Ls, LOB, Os	objective *wh-* pronoun	*whom, who, which, that* in non-subject position

（续表）

Tags		Used in	Meaning	Examples
WPS		Br, Ls	nominative *wh-* pronoun	*who*, *whoever*, *which*, *that* in subject position
	WPA	LOB	nominative *wh-*pronoun	*whosoever*
	WP	LOB	*wh-* pronoun neutral for case	*who*, *that* as relative pronoun
	WP	Os	interrogative *wh-* pronoun neutral for case	*who*, *whoever*
	WPR	Os	relative *wh-* pronoun neutral for case	*that*, *who*
	PNQS	La	*wh-* pronoun without *-ever*	*who*, *that*
	PNQVS	La	*wh-ever* pronoun	*whoever*
	WPO	LOB	objective *wh-* pronoun	*whom*, *whomsoever*
	WPO	Os	interrogative objective *wh-* pronoun	*whom*, *whomsoever*

（续表）

Tags		Used in	Meaning	Examples
	WPOR	Os	relative objective *wh-* pronoun	*whom*
	PNQO	La	objective *wh-* pronoun without *-ever*	*whom*
	PNQVO	La	objective *wh-ever* pronoun	*whomever*
WQL		Br, Ls, LOB, Os	*wh-* degree adverb	*how, however, in how green, however well*
	RGQ	La	*wh-* degree adverb without *-ever*	*how*
	RGQV	La	*wh-ever* degree adverb	*however*
WRB		Br, Ls, Os	non-degree *wh-* adverb	*when, how in how did it work*
	WRB	LOB	WQL and WRB	
	RRQ	La	non-degree *wh-* adverb without *-ever*	*where, when, why, how*

（续表）

Tags		Used in	Meaning	Examples
	RRQV	La	non-degree *wh-ever* adverb	*wherever, whenever, however*
	W	Go	all Brown W ... tags	

（注：上表根据 Garside, R. et al. 1987: 165 – 178 中 Appendix B 改编）



附录 3

汉英术语对照表

半固定词组	semi-fixed phrase
包含符	cover symbol
褒义词	ameliorative terms
本族语语料库	native language corpus
比较语料库	comparative corpus
比较语言学研究	comparative linguistic studies
比例抽样	proportionate sampling
贬义词	derogatory terms
标识	annotation
标注文件	annoted files
标注语料库	annoted corpus
标准化类符形符比	standardized type-token ratio
伯明翰英语文汇	Birmingham Collection of English Texts
参照语料库	reference corpus
超文本	hyper text
成语	idiom
词丛	word cluster
词丛表	word cluster table
词汇密度	lexical density

词块	lexical chunks
词类码	word class tag
词类码对子	pairs of word class tag
词频	word frequency
词频统计	word frequency statistics
词群	word groups
词图	plot
词项	lexical item
词项—语境方法	item-environment approach
词语词	lexical words
词语搭配	collocation
词语的决定作用	lexical determination
词语化	lexicalization
词语集	lexical set
词语结构	lexical structure
词语片语	lexical phrase
词语驱动	lexis-driven
词语型式	lexical pattern
词语行为	lexical behaviour
词语序列	lexical sequence
词语因素	lexical factors
词长	word length
词组	phrase
次技术词	sub-technical words
搭配词	collocates
搭配范围	collocational range
搭配分析	collocational analysis
搭配和谐	collocational harmony
搭配伙伴	collocational partner
搭配类	collocational class

搭配力	collocability
搭配强度	collocational strength
大纲设计	syllabus design
代表性	representativeness
数据驱动学习	Data-driven Learning
地道性	idiomaticity
地道语言使用	idiomatic language use
递归性	recursiveness
地域变体	regional variety
典型性	typicality
典型意义	typical meaning
动者格动词	ergative verb
对数或然率	log likelihood
多媒体计算机辅助语言学习	multi-media computer-assisted language learning (MCALL)
多特征	multi-feature
多维度	multi-dimension
反拨效应	backwash effect
反语	irony
方式准则	maxim of manner
方言研究	dialectal studies
非标记性	unmarkedness
非连续性搭配	discontinuous collocation
分层抽样	stratified sampling
峰态	peak
峰值	peak value
覆盖率	coverage
赋码	tagging
负偏态	negative skew
负迁移	negative transfer

负特征	negative feature
概念意义	ideational meaning
高频词	high frequency word
个人用语特点	idiolect
功能	function
功能词	function words
公认的语法	consensus grammar
公式块	formulaic chunks
共时语料库	synchronic corpus
共选	co-selection
观测频数	observed frequency
关键词	key words
关联准则	maxim of relevance
惯例化	institutionalization
惯例化搭配	institutionalized collocation
规避策略	avoidance strategy
合乎语法的搭配	well-formed collocations
核心搭配	core collocation
后编辑	post editing
话语分析	discourse analysis
话语社团	discourse community
会话合作原则	cooperative principles
混合语义韵	mixed prosody
机读电子文本	machine-readable electronic text
机读文本	machine-readable text
机读语料库	machine-readable corpus
积极语义韵	positive prosody
基子语料库	corpus-based
基于语料库的方法	corpus-based approach
基于语料库的研究	corpus-based study

技术词	technical words
技术文献	technical literature
计算语言学	computational linguistics
监控语料库	monitor corpus
节点词	node word
结构可变性	structural mutability
解决歧义码	tag disambiguation
句法	syntax
句法分析	parsing
句法片段	syntactic segment
聚合轴	paradigmatic axis
距位	span position
句长标准差	standard deviation of sentence length
段长标准差	standard deviation of paragraph length
角色关系	role relationship
科技术语	technical terms
可靠性	reliability
课程设置	course design
课件	courseware
口语语料库	spoken corpus
夸张	exaggeration
跨距	span
扩展搭配	extended collocation
扩展语境	extended context
类符	type
类符形符比	type/ token ratio
类联接	colligation
历时语料库	diachronic corpus
联想词汇	associative words

连续体	continuum
连续文本	running text
量化分析	quantitative analysis
名词化	nominalization
目的语	target language
内省	introspection
内省研究	introspective studies
能指	signifier
偶然共现词	casual collocating words
配价	valency
配码	tag assigning
篇章意义	textual meaning
频率	frequency
平均词长	average word length
平均段落长	average paragraph length
平均句长	average sentence length
平行语料库	parallel corpus
期望频数	expected frequency
起点频数	threshold frequency
权重	weighting
人工赋码	mannual tagging
人际意义	interpersonal meaning
任意性	arbitrariness
生成(JDEST)	generation
生语料库	raw corpus
施动者	agent
实词	content words
时间变体	diachronic variety
实时方法	real time approach
实证主义	empiricism

使用领域	area of use
受动者	patient
书面语语料库	written corpus
数据驱动	data-driven
数量准则	maxim of quantity
双语语料库	bilingual corpus
搜索词遮蔽	search word zapping
随机抽样	random sampling
所指	signified
通用词汇表	General Service List
通用语料库	general corpus
同形异义	homonymy
统计检验	statistical test
T-分值	T score
亡喻	dead metaphor
微型语境	mini-context
维度分数	dimensional score
文类	genre
习语赋码	idiom tagging
显著搭配	significant collocation
相对频数	relative frequency
相关矩阵	correlational matrix
相关系数	coefficient
相关性	relevance
相互吸引力	mutual attraction
相互信息值	MI value
相互预见	mutual prediction
消极语义韵	negative prosody
小文本	mini-text
效度	validity

心理词符	mental lexicon
心理因素	psychological factors
心灵主义	mentalism
新颖性	novelty
信道	channel
信息检索	information retrieval
信息焦点	information focus
形符	token
形式—意义配对	form-meaning pairing
形态分析	morphological analysis
行为主义	behaviorism
修辞格	figures of speech
修辞性搭配	figurative collocation
虚化词项	delexicalized word
学术英语	English for academic purposes
学习者语料库	learner corpus
学习者语言	learner language
言语识别	speech recognition
验证数据	attested data
样本	sample
一词多义	polysemy
一般搭配	general collocations
异常搭配	unusual collocations
义项	sense
意义	meaning
意义潜势	meaning potential
意义实现	meaning realization
因素分析	factor analysis
因循性	conventionality
隐喻	metaphor

英语文库	Bank of English
有限组合	restricted combination
右搭配词	right collocates
语步	move
语法词	grammatical words
语法的限制作用	syntactic restrictions
语法因素	grammatical factors
语阶	steps
语境	context of situation
语料档案	archives
语料库驱动	corpus-driven
语料库引擎	corpus engine
语篇连接语	discourse connectives
语体	style
语言变体	language variety
语言交际者	language communicators
语言经验	language experience
语言能力	linguistic competence
语言迁移	language transfer
语言社团	speech community
语言实体	linguistic entity
语言习得研究	language acquisition study
语言运用	linguistic performance
语义范畴	semantic category
语义分析	semantic analysis
语义和谐	semantic harmonization
语义明晰性	semantic transparency
语义韵	semantic prosody
语义韵冲突	prosodic clash
语音分析	phonetic analysis

语音合成	phonetic synthesis
语音识别	phonetic recognition
语音转写	phonetic transcription
语用错误	pragmatic error
语域	register
预编辑	pre-editing
预加工	pre-processing
原文本	raw text
原形还原	lemmatization
韵律	prosody
在线查询	on-line consultancy
在线语料库	on-line corpus
粘着性搭配	bound collocations
真实语言数据	authentic language data
正迁移	positive transfer
正态	normal
正特征	positive feature
正字法	orthography
正字分析	orthographic analysis
知识表述	representation of knowledge
直觉数据	intuitive data
质量准则	maxim of quality
制约语法	constraint grammar
中国学习者英语语料库	Chinese learners English Corpus
中间语	interlanguage
中间语对比分析	contrastive interlanguage analysis
中性语义韵	neutral prosody
主要关键词	key key word
专业文本	specialized text
专业性搭配	specialized collocation

专用语料库	specialized corpus
转写	transcription
自动对列	automatic alignment
自动赋码	automatic tagging
自然发生数据	naturally occurring data
自然性	naturalness
自由组合	free combinations
自由作文	free composition
组合关系	syntagmatic relation
组合选择	syntagmatic choice
组合因素	syntagmatic factors
组合轴	syntagmatic axis
左搭配词	left collocates
Z-分值	Z score
词典编纂	lexicography
词语索引	concordance
概率	probability
概率信息	probabilistic information
候选码	candidate tag
话语片段	discourse segment
或然语法	likelihood grammar
机器翻译	machine translation
计算机科学	computer science
计算语言学	computational linguistics
逻辑语义	logic semantics
内省法	introspection
篇章语言学	text linguistics
认知语言学	cognitive linguistics
文本	text
习语赋码	idiom tagging

习语码	idiom tag
相邻码渡越概率	transition probability
心灵主义	mentalism
形式语法	formal grammar
学术用途英语	English for Academic Purposes (EAP)
引出法	elicitation
应用语言学	applied linguistics
英语用法调查	Survey of English Usage
语感	intuitions
语料库	corpus
语料库语言学	corpus linguistics
语言工程	language engineering
语言能力	linguistic competence
语言运用	linguistic performance
语言哲学	language philosophy
专门用途英语	English for Specific Purposes (ESP)
转换生成语法	transformational grammar
自然语言处理	natural language processing

附录 4

英汉术语对照表

agent	施动者
ameliorative terms	褒义词
annotation	标识
annotated corpus	标注语料库
annotated files	标注文件
applied linguistics	应用语言学
arbitrariness	任意性
archives	语料档案
area of use	使用领域
associative words	联想词汇
attested data	验证数据
authentic language data	真实语言数据
automatic alignment	自动对列
automatic tagging	自动赋码
average paragraph length	平均段落长
average sentence length	平均句长
average word length	平均词长
avoidance strategy	规避策略
backwash effect	反拨效应

Bank of English	英语文库
behaviorism	行为主义
bilingual corpus	双语语料库
Birmingham Collection of English Texts	伯明翰英语文汇
bound collocations	粘着性搭配
candidate tag	候选码
casual collocating words	偶然共现词
channel	信道
Chinese Learners' English Corpus	中国学习者英语语料库
coefficient	相关系数
cognitive linguistics	认知语言学
colligation	类联接
collocates	搭配词
collocation	词语搭配
collocational analysis	搭配分析
collocational class	搭配类
collocational partner	搭配伙伴
collocational range	搭配范围
collocational strength	搭配强度
collocability	搭配力
collocational harmony	搭配和谐
comparative corpus	比较语料库
comparative linguistic studies	比较语言学研究
computational linguistics	计算语言学
computer science	计算机科学
concordance	词语索引
consensus grammar	公认的语法
constraint grammar	制约语法
content words	实词

context, context of situation	语境
continuum	连续体
contrastive interlanguage analysis	中间语对比分析
conventionality	因循性
conversational cooperative principles	会话合作原则
core collocation	核心搭配
corpus-driven	语料库驱动
corpus engine	语料库引擎
corpus linguistics	语料库语言学
corpus	语料库
corpus-based approach	基于语料库的方法
corpus-based study	基于语料库的研究
corpus-based	基于语料库
correlational matrix	相关矩阵
co-selection	共选
course design	课程设置
courseware	课件
cover symbol	包含符
coverage	覆盖率
Data-driven Learning	数据驱动学习
data-driven	数据驱动
dead metaphor	亡喻
delexicalized word	虚化词项
derogatory terms	贬义词
diachronic variety	时间变体
diachronic corpus	历时语料库
dialectal studies	方言研究
dimensional score	维度分数
discontinuous collocation	非连续性搭配

discourse analysis	话语分析
discourse community	话语社团
discourse connectives	语篇连接语
discourse segment	话语片段
elicitation	诱导法
empiricism	实证主义
English for Academic Purposes	学术用途英语
English for Specific Purposes	专门用途英语
ergative verb	动者格动词
exaggeration	夸张
expected frequency	期望频数
extended collocation	扩展搭配
extended context	扩展语境
factor analysis	因素分析
figurative collocation	修辞性搭配
figures of speech	修辞格
formal grammar	形式语法
form-meaning pairing	形式—意义配对
formulaic chunks	公式块
free combinations	自由组合
free composition	自由作文
frequency	频率
function words	功能词
function	功能
general collocations	一般搭配
general corpus	通用语料库
General Service List	通用词汇表
generation	生成
genre	文类
grammatical factors	语法因素

grammatical words	语法词
high frequency word	高频词
homonymy	同形异义
hyper text	超文本
ideational meaning	概念意义
idiolect	个人用语特点
idiom tagging	习语赋码
idiom	成语
idiomatic language use	地道语言使用
idiomaticity	地道性
idiom tag	习语码
information focus	信息焦点
information retrieval	信息检索
institutionalization	惯例化
institutionalized collocation	惯例化搭配
interlanguage	中间语
interpersonal meaning	人际意义
introspection	内省(法)
introspective studies	内省研究
intuitions	语感
intuitive data	直觉数据
irony	反语
item-environment approach	词项—语境方法
JDEST	上海交通大学科技英语语料库
key key word	主要关键词
key words	关键词
language acquisition study	语言习得研究
language communicators	语言交际者
language engineering	语言工程

language experience	语言经验
language philosophy	语言哲学
language transfer	语言迁移
language variety	语言变体
learner corpus	学习者语料库
learner language	学习者语言
left collocates	左搭配词
lemmatization	原形还原;削尾处理
lexical behaviour	词语行为
lexical chunks	词块
lexical density	词汇密度
lexical determination	词语的决定作用
lexical factors	词语因素
lexical item	词项
lexical pattern	词语型式
lexical phrase	词语片语
lexical sequence	词语序列
lexical set	词语集
lexical structure	词语结构
lexical words	词语词
lexicalization	词语化
lexicography	词典编纂
lexis-driven	词语驱动
likelihood grammar	或然语法
linguistic competence	语言能力
linguistic entity	语言实体
linguistic performance	语言运用
log likelihood	对数或然率
logic semantics	逻辑语义
machine translation	机器翻译

machine-readable corpus	机读语料库
machine-readable electronic text	机读电子文本
machine-readable text	机读文本
mannual tagging	人工赋码
maxim of manner	方式准则
maxim of quality	质量准则
maxim of quantity	数量准则
maxim of relevance	关联准则
meaning potential	意义潜势
meaning realization	意义实现
meaning	意义
mental lexicon	心理词符
mentalism	心灵主义
metaphor	隐喻
MI value	相互信息值
mini-context	微型语境
mini-text	小文本
mixed prosody	混合语义韵
monitor corpus	监控语料库
morphological analysis	形态分析
move	语步
multi-dimension	多维度
multi-feature	多特征
multi-media computer-assisted language learning	多媒体计算机辅助语言学习
mutual attraction	相互吸引力
mutual prediction	相互预见
native language corpus	本族语语料库
natural language processing	自然语言处理
naturally occurring data	自然发生数据

naturalness	自然性
negative feature	负特征
negative prosody	消极语义韵
negative skew	负偏态
negative transfer	负迁移
neutral prosody	中性语义韵
node word	节点词
normal	正态
nominalization	名词化
novelty	新颖性
observed frequency	观测频数
on-line consultancy	在线查询
on-line corpus	在线语料库
orthographic analysis	正字分析
orthography	正字法
pairs of word class tag	词类码对子
paradigmatic axis	聚合轴
parallel corpus	平行语料库
parsing	句法分析
patient	受动者
peak value	峰值
peak	峰态
phonetic analysis	语音分析
phonetic recognition	语音识别
phonetic synthesis	语音合成
phonetic transcription	语音转写
phrase	词组
plot	词图
polysemy	一词多义
positive feature	正特征

positive prosody	积极语义韵
positive transfer	正迁移
post editing	后编辑
pragmatic error	语用错误
pre-editing	预编辑
pre-processing	预加工
probability	概率
probabilistic information	概率信息
proportionate sampling	比例抽样
prosodic clash	语义韵冲突
prosody	韵律
psychological factors	心理因素
quantitative analysis	量化分析
random sampling	随机抽样
raw corpus	生语料库
raw text	原文本
real time approach	实时方法
recursiveness	递归性
reference corpus	参照语料库
regional variety	地域变体
register	语域
relative frequency	相对频数
relevance	相关性
reliability	可靠性
representation of knowledge	知识表述
representativeness	代表性
restricted combination	有限组合
right collocates	右搭配词
role relationship	角色关系
running text	连续文本

sample	样本
search word zapping	搜索词遮蔽
semantic analysis	语义分析
semantic category	语义范畴
semantic harmonization	语义和谐
semantic prosody	语义韵
semantic transparency	语义明晰性
semi-fixed phrase	半固定词组
sense	义项
significant collocation	显著搭配
signified	所指
signifier	能指
span position	距位
span	跨距
specialized collocation	专业性搭配
specialized corpus	专用语料库
specialized text	专业文本
speech community	语言社团
speech recognition	言语识别
speech synthesis	语言合成
spoken corpus	口语语料库
standard deviation of paragraph length	段长标准差
standard deviation of sentence length	句长标准差
standardized type-token ratio	标准化类符形符比
statistical test	统计检验
steps	语阶
stratified sampling	分层抽样
structural mutability	结构可变性

style	语体
sub-technical words	次技术词
Survey of English Usage	英语用法调查
syllabus design	大纲设计
synchronic corpus	共时语料库
syntactic restrictions	语法的限制作用
syntactic segment	句法片段
syntagmatic axis	组合轴
syntagmatic choice	组合选择
syntagmatic factors	组合因素
syntagmatic relation	组合关系
syntax	句法
T score	T－分值
tag assigning	配码
tag disambiguation	解决歧义码
tagging	赋码
target language	目的语
technical literature	技术文献
technical terms	科技术语
technical words	技术词
text	文本
text linguistics	篇章语言学
textual meaning	篇章意义
threshold frequency	起点频数
token	形符
transcription	转写
transformational grammar	转换生成语法
transition probability	相邻码渡越概率
type	类符
type/token ratio	类符形符比

typical meaning	典型意义
typicality	典型性
unmarkedness	非标记性
unusual collocations	异常搭配
valency	配价
validity	效度
weighting	权重
well-formed collocations	合乎语法的搭配
word class tag	词类码
word cluster	词丛
word cluster table	词丛表
word frequency	词频
word frequency statistics	词频统计
word groups	词群
word length	词长
Z score	Z-分值



附录 5

英汉人名对照表

Aarts, J. 2.6.1	阿茨
Altenberg, B. 2.6.1	阿尔滕伯格
Asao, Kojiro 2.6.1	朝雄幸次郎
Atkins, S. 2.2	阿特金斯
Benson, M. 5.4	本森
Biber, D. 7.4	贝博
Bloomfield, L 4.1	布隆菲尔德
Bradley, J. P 5.5	布拉德利
Chomsky, N. 1.1.1, 2.2, 4.1.1	乔姆斯基
Clarke, C. 1.4	克拉克
Clear, J. 2.2, 9	克里尔
Cowie, A. 3.1	考伊
Cruden, A. 2.4	克鲁登
Cruse, A. 6.1	柯鲁斯
Edwards, J. A. 2.3	爱德华兹
Ellis, R. 2.7	埃利斯

Enkvist, N. E. 7.2 恩克韦斯特

Firth, J. R. 2.4.1, 3.1 弗斯
Francis, G. 5.4, 3.3 弗兰西斯
Fries, C. C. 2.4 弗赖斯

Garside, R. 1.3 加塞德
Granger, S. 2.6.1 格兰杰
Greenbaum, S. 24.2 格林鲍姆
Grice, H. P. 8.3 格莱斯

Halliday, M. A. K. 2.4.1, 3.1, 4.3 韩礼德
Hanks, P. 3.1 汉克斯
Harris, Z. 2.4 哈里斯
Hunston, S. 5.4 亨斯顿

Jesperson, O. 2.4 叶斯珀森
Johansson, S. 2.4.2 乔恩森
Johnson, S. 2.4 约翰生
Jordan, R. 7.2 乔丹

Kaszubski, P. 2.6.1 卡斯苏布斯基
Kennedy, G. 2.4, 3.2 肯尼迪
Kettemann, 5.2 凯特曼
Kjellmer, G. 3.1 杰尔默
Kruisinga, E. 2.4 克鲁辛格
Kuçéra, H. 1.1.1, 2.4.2 库切拉

Lancashire, I. 5.5 兰开夏
Leech, G. 2.1, 2.2, 2.3, 2.4 里奇

Levinson, S. 8.3	莱文森
Lorenz, G. 2.6.1	洛伦茨
Louw, B. 8.1	洛悟
Mackin, R. 3.1	麦金
Martin, W. 3.2	马丁
McCarthy, M. 3.1	麦卡锡
McEnery, T. 2.2, 4.1.1	麦克艾纳利
McIntosh, A. 3.2	麦金托什
Miller, A. 8.3	米勒
Millic, L. 7.2	梅里克
Milton, J. 2.6.1	密尔顿
Mitchell, T. F. 3.1	米切尔
Miura, Shogo 2.6.1	宫昭二
Murison-Bowie, S. 2.3	穆里森伯韦
Nelson, F. 1.1.1, 2.4.2	纳尔逊
Ooi, V. 7.3	奥依
Pawley, A. 6.4	波利
Poutsma, H. 2.4	普茨默
Presutti, L. 5.5	普拉苏提
Quirk, R. 1.1.1, 5.2	夸克
Renouf, A. 2.2	赫努
Ringbom, H. 2.6.1	林伯姆
Rockwell, G. 5.5	洛克维尔
Rundell, M. 2.1	兰德尔

Russel, B. 1.3 罗素

Sapir, E. 1.3 萨丕尔
Saussure, de. 2.4.1 索绪尔
Sampson, G. 1. 3. 1 桑普森
Scott, M. 5.5 斯科特
Selinker, L. 5.4 塞林克
Sinclair, J. 1.2.3, 2.1, 2.3, 3.1, 4.1 辛克莱
Smadja, F. 5. 4 斯麦迦
Stairs, M. 5.5 斯德尔
Stubbs, M. 8.1 斯塔布斯
Svartvik, J. 1.2.2, 2.4.2 斯瓦特维克
Swales, J. 3.3, 6.4, 7.1 斯韦尔斯
Syder, H. 6.4 塞德

Thompson, 2.5 汤普森
Thorndike E. 1.4.1 桑代克
Tognini-Bonelli, E. 9.6 托格尼尼—布纳里
Traver, 2.4 特拉弗

West, M. 1.4.1, 4.1.1 威司特
Wilson, A. 2.4 威尔逊
Winter, W. 7.2 温特



参考书目

Aarts, J. 1996. *Advanced students of English found wanting* (Abstract). AILA.

Aitchison, J. 1987. *Words in the Mind: An Introduction to the Mental Lexicon.* Oxford: Basil Blackwell.

Aijmer, Karin. 1998. Modality in Swedish learners' written interlanguage. *The Learner Corpora and FL/SL Learning and Teaching Symposium, December* 14 – 16, 1998.

Altenberg, B. 1996. Exploring the Swedish Subcorpus of the International Corpus of English (Abstract). AILA

Atkins, S., Clear, J. & Ostler, N. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7: 1 – 16.

Baker, M., Francis, G & Tognini-Bonelli (eds). Text and Technology: in Hnour of Johm Sinclair, 157 – 176. Amsterdam: John Benjamins

Biber, D. 1988. *Variations Across Speech and Writing. Cambridge:* Cambridge University Press.

Biber, D. 1992. Using computer-based text corpora to analyze the referential strategies of spoken and written texts. In Svartvik (Ed) 1992: 213 – 252.

Chomsky, A. N. 1957. Extracts from *Syntactic Structures.* in Liu Runqing *et al,* (Eds) 1988, *Readings in Linguistics: Seventy-five years since Saussure.* Vol. 2: 86 – 120. Beijing: Cehui Press.

Chomsky, A. N. 1962. Paper given at the University of Taxes 1958, 3rd Taxes Conference on Problems of Linguistic Analysis in English, Austin: University of Taxes.

Chomsky, A. N. 1965. *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.

Church, K. W.& Hunks, P. 1990. Word Association Norms, Mutual Information & Lexicography. *Computational Linguistics* 16, 1. 22 – 29.

Clear, J. 1993. From Firth principles: computational tools for the study of collocation, In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds). *Text and Technology: In Honour of John Sinclair*, Amsterdam: John Benjamins

Corder, S. P. 1981. *Error Analysis and Interlanguage.* Oxford: Oxford University Press.

Cowie, A. and Mackin, R. 1975. *Oxford Dictionary of Current Idiomatic English*, Vol.1. London: Oxford University Press.

Cruse, D. A. 1989. *Lexical Semantics.* Cambridge: Cambridge University Press.

De Cock, S., Granger, S. Leech, G., and McEnery T. 1998. An automated approach to the phrasicon of EFL learners. In S. Granger (Ed). *Learner English on Computer.* 67 – 79. London: Longman.

Edwards, J. A. 1993. Survey of electronic corpora and related resources for language researchers. In Edwards & Lampert (Eds). *Talking Data: Transcription and Coding in Discourse Research.* 263 – 310. Hillsdale, N.J.: Erlbaum.

Enkvist, N. E. 1973. *Linguistic Stylistics.* Paris: Mouton & Co. Publisher.

Enkvist, N. E. 1964. *Linguistics and Style.* Paris: Mouton & Co. Publisher.

Firth, J. R. 1957. *Papers in linguistics 1934 – 1951*. London: Oxford University Press.

Firth, J. R. 1957b. A synopsis of linguistic theory, 1930 – 1955. in Liu Runqing *et al* (Eds). *Readings in Linguistics: Seventy-Five Years Since Saussure*. Vol. 1: 328 – 359. Beijing: Cehui Press.

Francis, G. 1993. A Corpus-Driven Approach to Grammar: Principles, Methods and Examples. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds). *Text and Technology: In Honour of John Sinclair*. 137 – 156. Amsterdam: John Benjamins.

Fries, C. C. 1952. *The Structure of English*. New York: Harcourt, Brace & Co.

Garside, R, Leech, G.& Sampson, G. (Eds). 1987. *The Computational Analysis of English. A Corpus-Based Approach*. London: Longman.

Garside, R. 1987. The CLAWS word-tagging system. in R. Garside, G. Leech & G. Sampson (Eds). *The Computational Analysis of English: A Corpus-Based Approach*. 30 – 41. London: Longman.

Garside, R. & Leech, G. 1987. The UCREL probabilistic parsing system. in R. Garside, G. Leech & G. Sampson (Eds). *The Computational Analysis of English: A Corpus-Based Approach*. 66 – 81. London: Longman.

Gitsaki, C. 1996. *Second Language Collocational Acquisition Research*. Unpublished PH. D. Thesis, Queensland University, Australia.

Granger, S. 1996. *The computerized learner corpus: a versatile source of data for SLA research* (Personal mail attachment).

Halliday, M. A. K. 1966. Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins

(Eds). *In Memory of J. R. Firth*. 148－162. London: Longman.

Halliday. M.A.K. 1976. Lexical relations. In Kress, C. (Ed). System and Function in Language. Oxford: Oxford University Press.

Halliday, M. A. K. 1978. *Language as Social Semiotic*. London: Adward Arnold.

Halliday, M. A. K. 1985. *An Introduction to Functional Grammar*. London: Adward Arnold.

Halliday, M. A. K. 1991. Corpus studies and probabilistic grammar. in K. Aijmer & B. Altenberg (Eds). *English Corpus Linguistics: Studies in Honor of Jan Svartvik*. 30－43. London: Longman.

Hanks, P. 1988. Definitions and Explanations. In J. Sinclair (Ed). *Looking Up*. 116－136. London: Collins COBUILD.

Howarth, P. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19/1: 24－44.

Hunston, S. and Francis, G. 1988. Verbs observed: A corpus-driven pedagogic grammar. *Applied Linguistics* 19/1: 45－72.

Jones, S. & Sinclair, J. 1974. *English Lexical Collocations: a Study in Computational Linguistics*. Cahiers de Lexicologie: 23(2). 15－61.

Jordan, R. R. 1997. *Enlgish for Academic Purposes*. Cambridge: Cambridge University Press.

Kaszubski, P. 1996. *Using authentic learner data to improve EFL writing textbooks — a heuristic approach* (Abstract). (personal communication).

Kennedy, G. 1998. *An introduction to corpus linguistics*. London: Longman.

Kennedy, G.D. 1990. Collocations: Where grammar and vocabulary teaching meet. In S. Anivan (Ed). *Language Teaching Methodology for the Nineties*. 215 – 229. Singapore: RELC.

Kjellmer, G. 1984. Some thoughts on collocational distinctiveness. In J. Aarts & W. Meijs (Eds). *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. 163 – 171. Costerus N.S. 45.

Kjellmer, G. 1987. Aspects of English collocations. In W. Meijs (Ed.). *Corpus Linguistics and Beyond*. 133 – 140. Amsterdam: Rodopi.

Kjellmer, G. 1991. A mint of phrases. In K. Aijmer & B. Altenberg (Eds). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. 111 – 127. London: Longman.

Krishnamurthy, R. 1987. The process of compilation. in J. M. Sinclair (Ed) *Looking Up*: 62 – 85. London: Collins *COBUILD*.

Leech, G. & C. Candlin (Eds). 1986. *Computers in English Language Teaching and Research*. London: Longman.

Leech, G. 1987. General introduction. In Garside et al (Eds) 1987: 1 – 15.

Leech, G. 1991. *Corpora, the Linguistics Encyclopedia*. K. Malmkjaer (Ed). 73 – 80. London: Routledge.

Leech, G. 1991. The state of the art in corpus linguistics. in K. Aijmer & B. Altenberg (Eds). *English Corpus Linguistics: Studies in Honor of Jan Svartvik*. 8 – 29. London: Longman.

Levinson, S. C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Liu Runqing, et. al (Eds). 1988. *Readings in Linguistics: Seventy-five Years Since Saussure*. Beijing: Cehui Press.

Louw, B. 1993. Irony in the Text or Insincerity in the Writer?

The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds). *Text and Technology: In Honour of John Sinclair*. 157 – 176. Amsterdam: John Benjamins.

Lyons, J. 1991. *Semantics* (Vol. 1). Cambridge: Cambridge University Press.

Martin, W. and P. J. G. van Sterkenburg. 1983. On the processing of a text corpus: from textual data to lexicographic information. In R. R. K. Hartmann (Ed). *Lexicography: Principles and Practice*. 77 – 87. London: Academic Press.

McCarthy, M. J. 1995. *Vocabulary*. Oxford: Oxford University Press.

McEnery, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McIntosh, A. 1967. Patterns and Ranges. In A. McIntosh and M. A. K. Halliday (Eds). *Patterns of Language: Papers in General, Descriptive and Applied Linguistics*. 181 – 199. Bloomington and London: Indiana University Press.

Mitchell, T. F. 1975. *Principles of Firthian Linguistics*. London: Longman.

Murison-Bowie, S. 1993. *MicroConcord Manual*. Oxford: Oxford University Press.

Nattinger, J. R. & DeCarrico, J. S. 1992. *Lexical Phrases and Language Learning*. Oxford: Oxford University Press.

Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi, Vincent B. Y. 1998. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Pawley, A., & Syder, H. 1983. Two puzzles for linguistic theory: Nativelike selection and Nativelike fluency. In J. C. Richard

& R.W. Schmidt (Eds). *Language and Communication*. 191 – 225. New York: Longman.

Renouf, A. 1986. *The Exploitation of a Computerised Corpus of English Text*. Actes du VIIIeme Colloque G. E. R. A. S. (Ed). Universite de Paris-Dauphine.

Renouf, A. 1987. Corpus development. In Sinclair, J. (Ed). 1987: 1 – 40.

Renouf, A. 1992. What do you think of that? A pilot study of the phraseology of the core words in English. In Leitner (Ed). *New Directions in English Language Corpora: Methodology, Results, Software Developments*. 301 – 317. Berlin: Mouton de Gruyter.

Renouf, A. and Sinclair. J. 1991. Collocational Frameworks in English. In Aijmer, K. & Altenberg, B. (Eds). *English Corpus Linguistics: Studies in Honour of Jan Svartvik* 128 – 143. London: Longman.

Ringbom, H. 1996. *Vocabulary Frequencies in Learner Language* (Abstract). AILA

Rundell, M. 1996. The corpus of the future, and the future of the corpus. Talk at Exeter, special conference on "New Trends in Reference Science".

Sampson, G. 1987. The grammatical database and parsing scheme. in R. Garside, G. Leech & G. Sampson (Eds). *The Computational Analysis of English: A Corpus-Based Approach*. 82 – 96. London: Longman.

Sapir, E. 1921. *Language*. Harcourt, Brace & World.

Selinker, L. 1972. Interlanguage. In J. C. Richards (Ed). 1984. *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman Group UK Limited.

Selinker, L. 1997. On the formation of interdialects: principles

from interlanguage. *Paper for webfeedback*: <*http://www.bbk.ac.uk/Departments/AppliedLinguistics/Larry.html*>.

Selinker, L. 1997. On the notion '1L competence' in early SLA research: an aid to understanding some baffling current issues. <*http://www.bbk.ac.uk/ Departments/ AppliedLinguistics/ Papers/ LarryCompAbs.html*>.

Selinker, L. et al. 1997. Multiple Language Acquisition: 'Damn it, why can't I keep these two languages apart?' In *Trilingual Language Acquisition*. Clevedon: Multilingual Matters.

Selinker, L. and Han, Zhaohong. 1997. *Fossilization: Status of the Concept*. London: Birkbeck College, University of London.

Selinker, L. and Han, Zhaohong. 1996. *Fossilization: What we Think we Know*. London: Birkbeck College, University of London.

Sinclair, J. 1966. Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M.A.K. Halliday & R. H. Robins (Eds). *In Memory of J. R. Firth*. 410 – 430. London: Longman.

Sinclair, J. 1986. Basic computer processing of long texts. In G. Leech & C. Candlin (Eds). 185 – 203.

Sinclair, J. 1987(Ed). *Collins COBUILD Dictionary of English Language*. London: Collins.

Sinclair, J. (Ed). 1987. *Looking up. London:* Collins COBUILD.

Sinclair, J. 1988. The Nature of Evidence. In J. Sinclair (Ed). *Looking Up*. 150 – 159. London: Collins COBUILD.

Sinclair, J. and Renouf, A. 1988. A lexical syllabus for language learning. in R. Carter & M. McCarthy (Eds). *Vocabulary and Language Teaching*. 140 – 160. London: Longman.

Sinclair, J. 1990 (Ed). *Collins COBUILD English Grammar.*

London: Collins COBUILD.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. 1991b (Ed). *Collins COBUILD Students' English Grammar.* London: Collins.

Sinclair, J. 1996. The Search for Units of Meaning, Textus IX. 75 – 106.

Sinclair, J. 1997. *Corpus Linguistics at the Millennium* (Personal communication).

Stubbs, M. 1996. *Text and Corpus Analysis.* Oxford: Blackwell Publishers.

Svartvik, J. (Ed). 1992. *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4 – 8 August 1991.* Berlin: Mouton de Gruyter.

Svartvik, J. 1996. Corpora are becoming mainstream. in J. Thomas & M. Short (Eds). *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech.* 3 – 13. London: Longman.

Swales, J. 1990. *Genre Analysis,* Cambridge: Cambridge University Press.

Wei, N. X. 1999. Towards Defining Collocations: A Practical Scheme for Study of Collocations in EAP Texts. Unpublished Ph. D. thesis. Shanghai Jiaotong University.

West, M. 1953. *A General Service List of English Words.* London: Longman.

Widdowson, H. G. 1989. Knowledge of language and ability for use. *Applied Linguistics.* 10: 128 – 137.

Willis, D. 1990. *The Lexical Syllabus. London:* Collins COBUILD.

Yang, H. 1986. A new technique for identifying scientific/tech-

nical terms and describing science texts. Literary and Linguistic Computing 1: 93 - 103.

Zhou,Qian and Yu,Shiwen. 1997. Annotating the Contemporary Chinese Corpus, *International Journal of Corpus Linguistics* Vol.2 (2). Amsterdam: John Benjamins.

陈建生 1998,语篇的自动词性赋码,《当代语言学》.75/1:17 - 28.

冯志伟 1996,《自然语言的计算机处理》.上海:上海外语教育出版社.

桂诗春、宁春岩 1997.《语言方法论》. 北京:外语教育与研究出版社.

雷秀云等 1999.《英语常用词 5000 + 5000》. 上海:上海交通大学出版社.

陆谷孙(主编)1989.《英汉大词典》. 上海:上海译文出版社.

濮建忠 2000. 中国学生英语动词语法和词汇型式使用特点初探.《现代外语》.1:24 - 24。

濮建忠、卫乃兴 2000. 词汇和语法.《解放军外国语学院学报》. 2 期:1 - 6.

钱瑗,1997. 对 COLLOCATION 的再认识.《外语教学与研究》. 111/3:43 - 47.

杨惠中 1997. 多媒体机助语言教学.《外语教学与研究》. 112/4:58 - 62.

参考使用的语料库及索引工具:

1. JDEST 语料库:上海交通大学语言文字工程研究所主建的学术英语语料库
2. LOB 语料库:Lancaster、Oslo 和 Bergen 大学合建的一般英国英语语料库
3. BROWN 语料库: Brown University 建立的一般美国英语语料库
4. COBUILD 语料库:Birmingham University 和 Collins 合建的一般英语语料库
5. Wordsmith Tools:Mike Scott, 1997
6. MicroConcord, Tom Johns & Mike Scott, 1996
7. TACT 2.1, Toronto University, 1984
8. Concordancer, R. J. C. Watt, 1999